

# On the Sample Complexity of Predictive Sparse Coding

Nishant A. Mehta\* and Alexander G. Gray

School of Computer Science  
College of Computing  
Georgia Institute of Technology  
Atlanta, Georgia

February 21, 2012

## Abstract

Predictive sparse coding algorithms recently have demonstrated impressive performance on a variety of supervised tasks, but they lack a learning theoretic analysis. We establish the first generalization bounds for predictive sparse coding. In the overcomplete dictionary learning setting, where the dictionary size  $k$  exceeds the dimensionality  $d$  of the data, we present an estimation error bound that is roughly  $O(\sqrt{dk/m} + \sqrt{s}/(\mu m))$ . In the infinite-dimensional setting, we show a dimension-free bound that is roughly  $O(k\sqrt{s}/(\mu\sqrt{m}))$ . The quantity  $\mu$  is a measure of the incoherence of the dictionary and  $s$  is the sparsity level. Both bounds are data-dependent, explicitly taking into account certain incoherence properties of the learned dictionary and the sparsity level of the codes learned on actual data.

**Keywords:** Statistical learning theory, luckiness, data-dependent complexity, dictionary learning, sparse coding, LASSO

## 1 Introduction

Learning architectures such as the support vector machine and other linear predictors enjoy strong theoretical properties and many empirical successes (Steinwart and Christmann, 2008; Kakade et al., 2009; Schölkopf and Smola, 2002), but a learning-theoretic study of many more complex learning architectures is lacking. Predictive methods based on *sparse coding* recently have emerged which simultaneously learn a data representation via a nonlinear encoding scheme and an estimator linear in that representation (Bradley and Bagnell, 2009; Mairal et al., 2010, 2009). A sparse coding representation  $z \in \mathbb{R}^k$  of a data point  $x \in \mathbb{R}^d$  is learned by representing  $x$  as a sparse linear combination of  $k$  atoms  $D_j \in \mathbb{R}^d$  of a dictionary  $D = (D_1, \dots, D_k) \in \mathbb{R}^{d \times k}$ . In the coding  $x \approx \sum_{j=1}^k z_j D_j$ , all but a few  $z_j$  are zero.

Predictive sparse coding methods such as Mairal et al. (2010)’s *task-driven dictionary learning* recently have achieved state-of-the-art results on many tasks, including the MNIST digits task. Whereas standard sparse coding minimizes an unsupervised, reconstructive  $\ell_2$  loss, predictive sparse coding seeks to minimize a supervised loss by optimizing a dictionary  $D$  and a predictor which takes encodings to  $D$  as input. It empirically has been observed that sparse coding can provide good abstraction by finding higher-level representations which are useful in predictive tasks (Yu et al., 2009). Intuitively, the power of prediction-driven dictionaries is that they pack more atoms in parts of the representational space where the prediction task is more difficult. Despite the empirical successes of predictive sparse coding methods, it is unknown how well they generalize in a theoretical sense.

In this work, we develop what to our knowledge are the first generalization error bounds for predictive sparse coding algorithms; in particular, we focus on  $\ell_1$ -regularized sparse coding. Maurer and Pontil (2008)

---

\*To whom correspondence should be addressed. Email: [niche@cc.gatech.edu](mailto:niche@cc.gatech.edu)

and Vainsencher et al. (2011) previously established generalization bounds for the classical, reconstructive sparse coding setting. There, the objective is to learn a representation of data by coding the data to a dictionary of atoms such that the data can be reconstructed with low error using only the coded data and the dictionary. Extending their analysis to the predictive setting introduces certain difficulties, the most salient being a seemingly unavoidable demand that the learned encoder be stable with respect to dictionary perturbations. Our analysis therefore is intimately tied to encoder stability properties.

The sparse encoder’s stability is characterized by properties central to sparse inference:

- The *sparsity level*  $s$  – the number of non-zeros in an encoding of a point.
- The  *$s$ -incoherence*  $\mu_s$  – the square of the minimum singular value among all  $s$ -column subdictionaries of  $D$ .
- The *coding margin* – a measure of a coordinate’s sign stability.

Since the sparsity level and the coding margin depend on the particular training sample, our learning bounds will depend on the maximum sparsity level and the minimum coding margin observed on the training sample. The need for encoder stability may be the price one pays for absconding from the (quite stable) prison of  $\ell_2$ -regularization. As in the luckiness frameworks of Shawe-Taylor et al. (1998) and Herbrich and Williamson (2002), our learning bounds depend on *a posteriori*-observable properties related to the training sample and the learned hypothesis; hence we will need to guarantee that these properties hold with high probability over a second, ghost sample.

We provide learning bounds for two core scenarios in sparse coding: the *overcomplete setting* of  $k \geq d$  and the infinite-dimensional setting where  $d \gg k$  or  $d$  is even infinite. Both bounds hold provided that  $m$  is larger than 3 times the inverse of the *permissible radius of perturbation*, a function linear in the coding margin, the sparsity level, the inverse of the  $s$ -incoherence, and the  $\ell_1$ -regularization parameter  $\lambda$ .

Our contributions are:

1. A forward stability result for the LASSO which holds under mild conditions. This result implies conditions for support preservation under dictionary perturbations.
2. In the overcomplete setting, a learning bound on the estimation error for predictive sparse coding that is essentially of order  $\sqrt{\frac{dk}{m}} + \frac{\sqrt{s}}{\mu_s m}$  (Corollary 1).
3. In the infinite-dimensional setting, a learning bound on the estimation error for predictive sparse coding that is *independent* of the dimension of the data; the bound is essentially of order  $\frac{k\sqrt{s}}{\mu_{2s}\sqrt{m}}$  (Corollary 2).

Whereas in the overcomplete setting the bound depends on the observed sparsity level and the coding margin on the training sample, in the infinite-dimensional setting the bound depends on these quantities as measured on both the training sample and a second, unlabeled sample not used for training. Both learning bounds contain a factor of  $k$ , and the optimal order of  $k$  is unknown — in both the predictive setting and the reconstructive setting. In addition to providing guarantees in the case of simultaneous dictionary and linear estimator learning (true predictive sparse coding), the presented bounds are quite general in that they also apply to the heuristic, two-stage algorithm where one is given a labeled training sample, learns a dictionary reconstructively (not using the labels) to fix the sparse codes, and then learns a linear estimator using the same training sample but with the labels as well.

The next section introduces the reconstructive and predictive forms of sparse coding. Section 3 sets up notation and presents two results: the Sparse Coding Stability Theorem (Theorem 1) and a useful symmetrization lemma. The overcomplete and infinite-dimensional settings are covered in Sections 4 and 5 respectively. In Section 6 we compare the results to recent learning bounds for unsupervised sparse coding. To allow the paper to flow smoothly, we provide proof sketches in the main paper and leave detailed proofs to the appendix.

## 2 Sparse coding, reconstructive and predictive

Suppose points  $x_1, \dots, x_m$  have been drawn from a probability measure  $P_X$  over  $B_{\mathbb{R}^d}$ , the unit ball in  $\mathbb{R}^d$ . The sparse coding problem is to represent each point  $x_i$  as a sparse linear combination of  $k$  basis vectors, or *atoms*  $D_1, \dots, D_k$ . The atoms form the columns of a *dictionary*  $D \in \mathcal{D}$ , where  $\mathcal{D}$  is a space of dictionaries  $\mathcal{D} := (B_{\mathbb{R}^d})^k$  and  $D_i = (D_i^1, \dots, D_i^d)^T$ . We will use this definition of  $\mathcal{D}$  from here on out. In this work, we use  $rB_{\mathbb{R}^d}$  to denote the origin-centered ball in  $\mathbb{R}^d$  scaled to radius  $r$ .

It will be useful to frame sparse coding in terms of an encoder  $\varphi_D$ :

$$\varphi_D(x) := \arg \min_z \|x - Dz\|_2^2 + \xi(z), \quad (1)$$

where  $\xi(\cdot)$  is a sparsity-inducing regularizer, or to be colorfully concise, a *sparsifier*.

**Reconstructive sparse coding.** The reconstructive sparse coding objective is

$$\min_{D \in \mathcal{D}} \mathbb{E}_{x \sim P_X} \|x - D\varphi_D(x)\|_2^2 + \xi(\varphi_D(x)),$$

where  $P_X$  is a probability measure on  $B_{\mathbb{R}^d}$ . Generalization bounds for the empirical risk minimization (ERM) variant of this objective have been established: [Maurer and Pontil \(2008\)](#) showed an  $O(k/\sqrt{m})$  bound that is *independent* of the dimension  $d$ ; this is useful when  $d \gg k$ , as in general Hilbert spaces. [Vainsencher et al. \(2011\)](#) handled the overcomplete setting, producing a bound that is  $O(\sqrt{dk}/\sqrt{m})$  as well as fast rates of  $O(dk/m)$ .

**Predictive sparse coding.** As in [Mairal et al. \(2010\)](#), the predictive setup minimizes a supervised loss with respect to a representation and an estimator linear in the representation. Define a space of linear hypotheses  $\mathcal{W} := rB_{\mathbb{R}^k}$ , for  $r > 0$ . Given a probability measure  $P$  on  $B_{\mathbb{R}^d} \times \mathcal{Y}$ , with  $\mathcal{Y} \subset \mathbb{R}$  in regression and  $\mathcal{Y} = \{-1, 1\}$  in binary classification, the predictive sparse coding objective is

$$\min_{D \in \mathcal{D}, w \in \mathcal{W}} \mathbb{E}_{(x,y) \sim P} \phi(y, \langle w, \varphi_D(x) \rangle) + \Theta(w), \quad (2)$$

where  $\Theta(\cdot)$  is a regularizer often taken to be proportional to the squared  $\ell_2$ -norm.

Other formulations exist, some of which wield a separate dictionary for each class ([Mairal et al., 2008](#)); we do not consider such formulations here for multiple reasons. First, they do not naturally extend to regression problems. Also, when the number of classes is large, both the computational and sample complexities of learning a dictionary for each class becomes prohibitive.

**Encoder stability.** The choice of the sparsifier  $\xi$  seems to be pivotal both from an empirical perspective and a theoretical one. [Bradley and Bagnell \(2009\)](#) used a differentiable *approximate* sparsifier based on the Kullback-Leibler divergence. The sparsifier is approximate because true sparsity does not result, although it encourages low  $\ell_1$  norms; nevertheless, true sparsity is a necessity in certain applications and also aids in some theoretical arguments. The most popular sparsifier is  $\xi(\cdot) = \lambda \|\cdot\|_1$ . Notably,  $\|\cdot\|_1$  is the tightest convex lower bound for the  $\ell_0$  “norm”:  $|\{i : x_i \neq 0\}|$ .

From a stability perspective the  $\ell_1$  regularizer regrettably is not well-behaved in general. Indeed, from the lack of strict convexity it is evident that each  $x$  need not have a unique image under  $\varphi_D$ . It also is unclear how to analyze the class of mappings  $\varphi_D$ , parameterized by  $D$ , if the map changes drastically under small perturbations to  $D$ . Hence, we will begin by establishing sufficient conditions under which  $\varphi_D$  is stable under perturbations to  $D$ .

In this work, we analyze the ERM variant of (2) with  $\Theta(w) = \frac{1}{r} \|w\|_2^2$ :

$$\min_{D \in \mathcal{D}, w \in \mathcal{W}} \frac{1}{m} \sum_{i=1}^m \phi(y_i, \langle w, \varphi_D(x_i) \rangle) + \frac{1}{r} \|w\|_2^2. \quad (3)$$

Since this objective is not convex, we will present bounds on the estimation error that hold uniformly over certain random subclasses of hypotheses.

### 3 Definitions, notation, and foundational results

Some definitions and notation will be useful. Let  $\mathbf{z}$  be an iid random sample of  $m$  points, where each  $\mathbf{z}_i = (x_i, y_i)$ ,  $x_i \in B_{\mathbb{R}^d}$  and  $y_i \in \mathcal{Y}$ . Let the space of targets,  $\mathcal{Y}$ , be a bounded subset of  $\mathbb{R}$  (regression) or  $\{-1, 1\}$  (binary classification). Let  $\mathbf{x}''$  be a second, unlabeled iid sample of  $m$  points  $x_1'', \dots, x_m''$ ; this sample will be used only in the infinite-dimensional setting. Also, let  $\mathbf{z}'$  be an iid ghost sample of  $m$  points with the same distribution as  $\mathbf{z}$ .

A predictive sparse coding hypothesis function  $f$  is fully specified by a choice  $(D, \mathbf{w}) \in \mathcal{D} \times \mathcal{W}$ , yielding  $f(\mathbf{x}) = \langle \mathbf{w}, \varphi_D(\mathbf{x}) \rangle$ . We often identify  $f$  using the notation  $f = (D, \mathbf{w})$ . The function class  $\mathcal{F}$  is the set of such hypotheses. When provided a training sample  $\mathbf{z}$ , the hypothesis returned by the learner will be referred to as  $\hat{f}$ . Note that  $\hat{f}$  is random, but  $\hat{f}$  becomes a fixed function upon conditioning on  $\mathbf{z}$ .

Throughout this work,  $\phi : \mathcal{Y} \times \mathbb{R} \rightarrow [0, b]$  will be a bounded loss function ( $b > 0$ ) that is  $L$ -Lipschitz in its second argument. For  $f \in \mathcal{F}$ , define  $f_\phi : \mathcal{Y} \times \mathbb{R}^d \rightarrow \mathbb{R}_+$  as the loss-composed function  $f_\phi(y, \mathbf{x}) \mapsto \phi(y, f(\mathbf{x}))$ . Let  $\phi \circ \mathcal{F}$  be the class of such functions induced by the choice of  $\mathcal{F}$  and  $\phi$ . We use the notation

$$\mathbb{P} f = \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}} f(\mathbf{x}) \qquad \mathbb{P} f_\phi = \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}} \phi(y, f(\mathbf{x}))$$

for the expectation operator  $\mathbb{P}$  and the notation

$$\mathbb{P}_{\mathbf{z}} f = \frac{1}{m} \sum_{i=1}^m f(\mathbf{x}_i) \qquad \mathbb{P}_{\mathbf{z}} f_\phi = \frac{1}{m} \sum_{i=1}^m \phi(y_i, f(\mathbf{x}_i))$$

for the empirical measure  $\mathbb{P}_{\mathbf{z}}$  associated with sample  $\mathbf{z}$ .

Define  $\text{LASSO}(\lambda, D, \mathbf{x})$  as the optimization problem

$$\text{LASSO}(\lambda, D, \mathbf{x}) \equiv \min_{\alpha \in \mathbb{R}^k} \|\mathbf{x} - D\alpha\|_2^2 + \lambda \|\alpha\|_1;$$

call the argmin  $\varphi_D(\mathbf{x}) = ((\varphi_D(\mathbf{x}))_1, \dots, (\varphi_D(\mathbf{x}))_k)^T$  ( $\lambda$  is fixed and hence withheld from the notation). Let  $\mathbb{R}_+$  be the set of positive reals and  $[n] := \{1, \dots, n\}$ , for  $n \in \mathbb{N}$ . For  $s \in [k]$ , any  $d \geq s$ , and  $D \in \mathcal{D}$ , define  $\mu_s(D)$  as the minimum squared singular value, taken across all matrices formed by choosing  $s$  distinct columns of  $D$ . We often use the overloaded definition  $\mu_s(f) := \mu_s(D)$  where  $f = (D, \mathbf{w})$ . The index set  $\mathcal{A}$  is the set of non-zero indices of  $\varphi_D(\mathbf{x})$ ; with some abuse of notation  $\mathcal{A}$  always is defined in terms of the most recently used dictionary  $D$  and data point  $\mathbf{x}$ . Index sets induce coordinate projections in the following way: if  $\mathbf{z} \in \mathbb{R}^k$ , then  $\mathbf{z}_{\mathcal{A}} = (z_{\mathcal{A}_1}, \dots, z_{\mathcal{A}_{|\mathcal{A}|}})$ . For  $\mathbf{t} \in \mathbb{R}^k$ , define  $\text{supp}(\mathbf{t}) := \{i \in [k] : t_i \neq 0\}$ .

Finally, an *epsilon-cover* refers to the concept of an  $\varepsilon$ -cover but not a specific cover. All epsilon-covers of spaces of dictionaries use the metric induced by the operator norm  $\|\cdot\|_2$ .

#### 3.1 Sparse coding stability

We begin with a fundamental forward stability result for the LASSO; we are not aware of any similar result in the literature.

**Theorem 1 (Sparse Coding Stability).** *Let  $\lambda > 0$ ,  $\mathbf{x} \in B_{\mathbb{R}^d}$ , and  $D, \tilde{D} \in (B_{\mathbb{R}^d})^k$  with  $\mu_s(D), \mu_s(\tilde{D}) \geq \mu$  for  $\mu > 0$ . If there are  $\tau > 0$  and  $s \in [k]$  such that:*

$$(i) \qquad \|\varphi_D(\mathbf{x})\|_0 \leq s, \tag{4}$$

$$(ii) \qquad |(\varphi_D(\mathbf{x}))_j| > \tau \text{ for all } j \in \mathcal{A}, \tag{5}$$

$$(iii) \qquad |\langle D_j, \mathbf{x} - D\varphi_D(\mathbf{x}) \rangle| < \lambda - \tau \text{ for all } j \notin \mathcal{A}; \tag{6}$$

and if

$$\|D - \tilde{D}\|_2 \leq \varepsilon \leq \frac{\tau\mu}{\frac{s+\mu}{\lambda} + \sqrt{s+\mu}}, \tag{7}$$

then

$$\text{supp}(\varphi_D(x)) = \text{supp}(\varphi_{\tilde{D}}(x))$$

and

$$\|\varphi_D(x) - \varphi_{\tilde{D}}(x)\|_2 \leq \frac{\varepsilon}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right).$$

Condition (6) means that atoms not used in the coding  $\varphi_D(x)$  cannot have too high absolute correlation with the residual  $x - D\varphi_D(x)$ . Note that the right-most term of (7) is the permissible radius of perturbation (PRP). In short, the theorem says that if  $\text{LASSO}(\lambda, D, x)$  admits a stable sparse solution, then a small perturbation to the dictionary will not change the support of the solution, and the perturbation to the solution will be bounded by a constant factor times the size of the perturbation (where the constant depends on a condition number to a restricted problem, the amount of  $\ell_1$ -regularization, and the sparsity level).

*Proof sketch:* Our strategy is to show that there is a unique solution to the perturbed problem, defined in terms of the optimality conditions of the LASSO (see conditions L1 and L2 of (Asif and Romberg, 2010)), and this solution has the same support as the solution to the original problem. As a result, the perturbed solution's proximity to the original solution is governed in part by a condition number  $\mu$  of a linear system of  $s$  variables. ■

### 3.2 Symmetrization by ghost sample for random subclasses

The next result is essentially due to Mendelson and Philips (2004); it applies symmetrization by a ghost sample for random subclasses. Our main departure is that we allow the random subclass to depend on a second, unlabeled sample  $\mathbf{x}''$ . The lemma will be applied to the overcomplete setting in the simpler form without  $\mathbf{x}''$  and to the infinite-dimensional setting in its full form.

**Lemma 1 (Symmetrization by Ghost Sample).** *Let  $\mathcal{F}(\mathbf{z}, \mathbf{x}'') \subset \mathcal{F}$  be a random subclass which can depend on both an labeled sample  $\mathbf{z}$  and an unlabeled sample  $\mathbf{x}''$ .*

*If  $m \geq \left(\frac{b}{\varepsilon}\right)^2$ , then*

$$\Pr_{\mathbf{z}\mathbf{x}''} \{ \exists f \in \mathcal{F}(\mathbf{z}, \mathbf{x}'') \text{ } P f_\Phi - P_{\mathbf{z}} f_\Phi \geq t \} \leq 2 \Pr_{\mathbf{z}\mathbf{z}'\mathbf{x}''} \left\{ \exists f \in \mathcal{F}(\mathbf{z}, \mathbf{x}'') \text{ } P_{\mathbf{z}'} f_\Phi - P_{\mathbf{z}} f_\Phi \geq \frac{t}{2} \right\}.$$

## 4 Overcomplete setting

Classically, the overcomplete setting is the *modus operandi* in sparse coding. Since  $k \geq d$  in this setting, an ideal learning bound has minimal dependence on  $k$ . We derive learning bounds with a square root dependence on  $k$ , below which the possibility of further improvement is open even in the reconstructive setting<sup>1</sup>. At a high level, our strategy for the overcomplete case learning bound is to construct an epsilon-cover over a subclass of the space of functions  $\mathcal{F} := \{f = (D, \mathbf{w}) : D \in \mathcal{D}, \mathbf{w} \in \mathcal{W}\}$  and to show that the metric entropy of this subclass is of order  $dk$ . The main difficulty is that an epsilon-cover over  $\mathcal{D}$  need not approximate  $\mathcal{F}$  to any degree, *unless* one has a notion of encoder stability (which we describe in this section). Our analysis effectively will be concerned only with a training sample and a ghost sample; hence, similar to the style of the luckiness framework of Shawe-Taylor et al. (1998), should we observe that the sufficient conditions for encoder stability hold true on the training sample, it is enough to guarantee that most points in a ghost sample also satisfy these conditions (possibly at a weaker level).

### 4.1 Useful conditions and subclasses

We first establish two important conditions. Letting  $f = (D, \mathbf{w}) \in \mathcal{D} \times \mathcal{W}$ , define

$$A_s(f, \mathbf{x}) := \left\{ \max_{x_i \in \mathbf{x}} \|\varphi_D(x_i)\|_0 \leq s \right\} \quad \text{and} \quad C_\tau(f, \mathbf{x}) := \left\{ \max_{x_i \in \mathbf{x}} \max_{j \in [k]} \psi_j(D, x_i) < \lambda - \tau \right\},$$

<sup>1</sup>The only works known to attack the reconstructive setting are Maurer and Pontil (2008) and Vainsencher et al. (2011).

where  $\psi_j(D, \mathbf{x}) := |\langle D_j, \mathbf{x} - D\varphi_D(\mathbf{x}) \rangle| - |(\varphi_D(\mathbf{x}))_j|$ . The first condition is critical as the learning bound will exploit the sparsity level over the training sample; the second condition can be motivated as follows. Fix some  $\mathbf{x} \in B_{\mathbb{R}^d}$  and assume  $C_\tau(f, \mathbf{x})$  is true. If  $j \in \mathcal{A}$ , the LASSO optimality condition L1 of [Asif and Romberg \(2010\)](#) implies that  $|\langle D_j, \mathbf{x} - D\varphi_D(\mathbf{x}) \rangle| = \lambda$ ; since  $\psi_j(D, \mathbf{x})$  is true, it follows that condition (5) holds. If  $j \notin \mathcal{A}$ , then since  $(\varphi_D(\mathbf{x}))_j = 0$  the condition  $\psi_j(D, \mathbf{x})$  is precisely the condition (6). Hence, the condition  $\psi_j(D, \mathbf{x}) < \lambda - \tau$  is the key encoder-stability inequality which will need to be enforced for each coordinate. Since the form of  $\psi_j$  is independent of  $j$ , all coordinates can be treated identically independent of their membership in  $\mathcal{A}$ .

For notational compactness, define  $\mathcal{I}_{s,\tau}(f, \mathbf{x}) = A_s(f, \mathbf{x}) \wedge C_\tau(f, \mathbf{x})$  and

$$\mathcal{I}_{\eta,s,\tau}(f, \mathbf{x}) = \left( \nexists \tilde{\mathbf{x}} \subset \mathbf{x} \mid \tilde{\mathbf{x}} \mid > \eta \wedge \forall \mathbf{x} \in \tilde{\mathbf{x}} \overline{A_s(f, \mathbf{x})} \vee \overline{C_\tau(f, \mathbf{x})} \right),$$

where  $\bar{E}$  is the negation of a boolean expression  $E$ .

Define the level  $\alpha$ -coding margin  $\tau_\alpha(D, \mathbf{x}) := \max\{\tau' > 0 : C_{\alpha\tau'}(D, \mathbf{x})\}$  for  $\alpha > 0$  and the coding margin  $\tau(D, \mathbf{x}) := \tau_1(D, \mathbf{x})$ ; clearly  $\tau_\alpha(D, \mathbf{x}) = \alpha^{-1}\tau(D, \mathbf{x})$ . For convenience we often make the first argument  $f$  rather than  $D$ , using the overloaded definition  $\tau(f, \mathbf{x}) := \tau(D, \mathbf{x})$  where  $f = (D, \mathbf{w})$ . Our bounds will require a crucial PRP-based condition that depends on both the learned dictionary and the training sample:

$$\tau(D, \mathbf{x}) \geq \iota(\lambda, \mu, \varepsilon, s) \quad \text{for } \iota(\lambda, \mu, \varepsilon, s) = 3\varepsilon \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{\mu} + \frac{1}{\lambda} + 1 \right).$$

For brevity we will refer to  $\iota$  with its parameters implicit; the dependence on  $\varepsilon, s, \lambda$ , and  $\mu$  will be a nonissue because we first develop bounds with all these quantities fixed *a priori*.

Finally, we will use the subclasses  $\mathcal{D}_\mu := \{D \in \mathcal{D} : \mu_s(D) \geq \mu\}$  and  $\mathcal{F}_\mu := \{f = (D, \mathbf{w}) \in \mathcal{F} : D \in \mathcal{D}_\mu\}$ , where  $\mu > 0$  is fixed.

## 4.2 Learning bound

The following proposition is a specialization of Lemma 1 with  $\mathbf{x}''$  taken as the empty set and the random subclass defined as  $\mathcal{F}(\mathbf{z}, \mathbf{x}'') := \{f \in \mathcal{F}_\mu : \mathcal{I}_{s,\iota}(f, \mathbf{x})\}$ .

**Proposition 1.** *If  $m \geq \left(\frac{b}{t}\right)^2$ , then*

$$\begin{aligned} & \Pr_{\mathbf{z}} \{ \exists f \in \mathcal{F}_\mu \mathcal{I}_{s,\iota}(f, \mathbf{x}) \wedge (\mathbf{P} f_\Phi - \mathbf{P}_{\mathbf{z}} f_\Phi > t) \} \\ & \leq 2 \underbrace{\Pr_{\mathbf{z}\mathbf{z}'} \{ \exists f \in \mathcal{F}_\mu \mathcal{I}_{s,\iota}(f, \mathbf{x}) \wedge (\mathbf{P}_{\mathbf{z}'} f_\Phi - \mathbf{P}_{\mathbf{z}} f_\Phi > t/2) \}}_{\text{event } J}. \end{aligned} \quad \blacksquare$$

Define  $Z$  as the event that there exists a hypothesis with stable codings on the original sample but more than  $\eta(m, d, k, \varepsilon, \delta)$  points with unstable codings on the ghost sample:

$$Z := \left\{ \mathbf{z}\mathbf{z}' : \exists f \in \mathcal{F}_\mu \mathcal{I}_{s,\iota}(f, \mathbf{x}) \wedge \left( \exists \tilde{\mathbf{x}} \subset \mathbf{x}' \mid \tilde{\mathbf{x}} \mid > \eta(m, d, k, \varepsilon, \delta) \wedge \forall \mathbf{x} \in \tilde{\mathbf{x}} \overline{\mathcal{I}_{s,\tau_3(f,\mathbf{x})}(f, \mathbf{x})} \right) \right\}.$$

We will show that  $\Pr(J)$  is small by use of the fact that

$$\Pr(J) = \Pr(J \cap \bar{Z}) + \Pr(J \cap Z) \leq \Pr(J \cap \bar{Z}) + \Pr(Z).$$

So far, our strategy is similar to the beginning of Shawe-Taylor et al.'s proof of the main luckiness framework learning bound (see [Shawe-Taylor et al., 1998](#), Theorem 5.22). We show that  $\Pr(Z)$  and  $\Pr(J \cap \bar{Z})$  are small in turn and then present the main learning bound.

The next lemma shadows [Shawe-Taylor et al. \(1998\)](#)'s notion of probable smoothness and vanquishes the key difficulty in bounding  $\Pr(Z)$ .

**Lemma 2 (Good Ghost).** Fix  $\mu, \lambda > 0$  and  $s \in [k]$ . Let  $\mathbf{x} \sim \mathcal{P}^m$  and  $\mathbf{x}' \sim \mathcal{P}^m$  be iid samples of  $m$  points each. Choose any  $D \in \mathcal{D}_\mu$  for which  $A_s(D, \mathbf{x})$ . With probability at least  $1 - \delta$  at least  $m - \eta(m, d, k, \varepsilon, \delta)$  points  $\tilde{\mathbf{x}} \subset \mathbf{x}'$  satisfy  $A_s(D, \tilde{\mathbf{x}})$  and  $C_{\tau_3(D, \mathbf{x})}(D, \tilde{\mathbf{x}})$ , for

$$\eta(m, d, k, \varepsilon, \delta) := 2dk \log \frac{96s}{\lambda \mu \tau(D, \mathbf{x})} + \log(2m + 1) + 2 \log \frac{2}{\delta}.$$

*Proof sketch:* The core idea of the proof is to show that if  $D$  satisfies encoder stability at level  $s$  and coding margin  $\tau$ , then there exists a representative  $D'$  in an  $\varepsilon$ -cover of  $\mathcal{D}_\mu$  — composed of balls with radius linear in  $\frac{1}{\tau}$  — which satisfies encoder stability at level  $s$  and a coding margin reduced by a constant factor. Next, a standard permutation argument (Vapnik and Chervonenkis, 1968, Proof of Theorem 2), along with a VC dimension argument to entertain all choices of the coding margin  $\tau$ , guarantees with high probability that the representative  $D'$  also satisfies the same stability properties on most of the ghost sample. Finally, we once again jump from  $D'$  to  $D$  to guarantee that, on this good part of the ghost sample where  $D'$  satisfies  $s$ -sparsity and slightly reduced coding margin,  $D$  also satisfies  $s$ -sparsity and some further reduced coding margin. ■

It remains to bound  $\Pr(J \cap \bar{Z})$ . We use the shorthand  $\eta = \eta(m, d, k, \varepsilon, \delta)$  for conciseness.

**Lemma 3 (Large Deviation on Good Ghost).** Let  $\varpi := t/2 - (2L\beta + \frac{b\eta}{m})$  and  $\beta := \varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right)$ . Then

$$\Pr(J \cap \bar{Z}) \leq \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)).$$

Equivalently, the difference between the loss on  $\mathbf{z}$  and the loss on  $\mathbf{z}'$  is greater than  $\varpi + 2L\beta + \frac{b\eta}{m}$  with probability at most  $\left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2))$ .

*Proof sketch:* From the definition of  $J \cap \bar{Z}$ , the entire training sample and all but  $\eta$  points of the ghost sample are “good” (i.e. have stable codings). Thus, there exists a representative  $f'$  in a product of  $\varepsilon$ -covers of  $\mathcal{D}$  and  $\mathcal{W}$  which closely approximates (with error at most  $2L\beta$ , from the Sparse Coding Stability Theorem (Theorem 1)) the function evaluation of all good points and suffers error at most  $\frac{b\eta}{m}$  on the bad points. Hence, for there to exist a large deviation  $t'$  between the training and ghost sample, there would need to be a deviation of at least  $t' - (2L\beta + \frac{b\eta}{m})$  for at least one hypothesis in the product of  $\varepsilon$ -covers. The result follows from Hoeffding’s inequality and the union bound. ■

A learning bound is within reach. Let  $\varepsilon = \frac{1}{m}$ ; then from some elementary manipulations:

**Theorem 2.** If  $m > \frac{1}{3\tau} \left( \frac{\frac{s}{\mu} + \sqrt{s}}{\mu} + 1 \right)$ , then for all  $f \in \mathcal{F}$  such that  $\mu_s(f) \geq \mu$ ,  $A_s(f, \mathbf{x})$ , and  $C_l(f, \mathbf{x})$ , with probability at least  $1 - \delta$  over  $\mathbf{z}$ :

$$\begin{aligned} (P - P_z)f_\Phi &\leq 2b \sqrt{\frac{2((d+1)k \log(8m) + k \log \frac{r}{2} + \log \frac{4}{\delta})}{m}} \\ &\quad + \frac{4L}{m} \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{2b}{m} \left( 2dk \log \frac{96s}{\lambda \mu \tau(f, \mathbf{x})} + \log(2m + 1) + 2 \log \frac{8}{\delta} \right). \end{aligned}$$

Now, we construct a bound for each choice of  $s$  and  $\mu$ . To each choice of  $s \in [k]$  assign prior probability  $\frac{1}{k}$ . To each choice of  $i \in \mathbb{N} \cup \{0\}$  for  $2^{-i} \leq \mu$  assign prior probability  $(i+1)^{-2}$ . For a given choice of  $s \in [k]$  and  $2^{-i} \leq \mu$  we use  $\delta(s, i) := \frac{6}{\pi^2} \frac{1}{(i+1)^2} \frac{1}{k} \delta$ . Note that  $\sum_{s=0}^k \sum_{i=0}^\infty \delta(s, i) = \delta$  since  $\sum_{i=1}^\infty \frac{1}{i^2} = \frac{\pi^2}{6}$ .

Define  $\nu(\mu) := \min\{1, 2^{\lfloor \log_2 \mu \rfloor}\}$  for  $\mu > 0$ . The following corollary is now proved.



**Corollary 1 (Overcomplete Learning Bound).** For any  $s \in [k]$ , for all  $f \in \mathcal{F}$  such that  $A_s(f, \mathbf{x})$ ,  $C_i(f, \mathbf{x})$ , and  $m > \frac{3}{\tau(f, \mathbf{x})} \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{v(\mu_s(f))} + \frac{1}{\lambda} + 1 \right)$ , with probability at least  $1 - \delta$  over  $\mathbf{z}$ :

$$\begin{aligned} (P - P_z)f_\phi &\leq 2b \sqrt{\frac{2 \left( (d+1)k \log(8m) + k \log \frac{r}{2} + \log \frac{2\pi^2 \left( \log_2 \frac{2}{v(\mu_s(f))} \right)^2 k}{3\delta} \right)}{m}} \\ &\quad + \frac{4L}{m} \left( \frac{r}{v(\mu_s(f))} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) \\ &\quad + \frac{2b}{m} \left( 2dk \log \frac{96s}{\lambda v(\mu_s(f)) \tau(f, \mathbf{x})} + \log(2m+1) + 2 \log \frac{4\pi^2 \left( \log_2 \frac{2}{v(\mu_s(f))} \right)^2 k}{3\delta} \right). \quad \blacksquare \end{aligned}$$

## 5 Infinite-dimensional setting

Often in learning problems, we first map the data implicitly to a space of very high dimension or even infinite dimension and use kernels for efficient computations. In these cases where  $d \gg k$  or  $d$  is infinite, it is unacceptable for any learning bound to exhibit dependence on  $d$ . Unfortunately, the strategy of the previous section breaks down in the infinite-dimensional setting because the straightforward construction of any epsilon-cover over the space of dictionaries had cardinality that depends on  $d$ . Even worse, epsilon-covers actually were used both to approximate the function class  $\mathcal{F}$  in  $\|\cdot\|_\infty$  norm and to guarantee that most points of the ghost sample are good provided that all points of the training sample were good (the Good Ghost Lemma (Lemma 2)).

These issues can be overcome by requiring an additional, unlabeled sample — a device often justified in supervised learning problems because unlabeled data may be inexpensive and yet quite helpful — and by switching to more sophisticated techniques based on conditional Rademacher and Gaussian averages. After learning a hypothesis  $\hat{f}$  from a predictive sparse coding algorithm, the sparsity level and coding margin are measured on a second, unlabeled sample  $\mathbf{x}''$  of  $m$  points<sup>2</sup>. Since this sample is independent of the choice of  $\hat{f}$ , it is possible to guarantee that all but a very small fraction ( $\frac{\eta}{m} = \frac{2 \log \frac{2}{\delta}}{m}$ ) of points of a ghost sample  $\mathbf{z}$  are good with probability  $1 - \delta$ . In the likely case of this good event, and for a fixed sample, we then consider all possible choices of a set of  $\eta$  bad indices in the ghost sample; each of the  $\binom{m}{\eta}$  cases corresponds to a subclass of functions. We then approximate each subclass by a special  $\varepsilon$ -cover that is a disjoint union of a finite number of special subclasses; for each of these smaller subclasses, we bound the conditional Rademacher average by exploiting a sparsity property.

### 5.1 Symmetrization and decomposition

Let  $\mu^* \in \mathbb{R}_+^2$  and define  $\mathcal{F}_{\mu^*} = \{f \in \mathcal{F} : (\mu_s(f) \geq \mu_s^*) \wedge (\mu_{2s}(f) \geq \mu_{2s}^*)\}$ . The next result is immediate from Lemma 1 where  $\mathcal{F}(\mathbf{z}, \mathbf{x}'') := \{\hat{f}\} \cap \{f \in \mathcal{F}_\mu : \gamma_{s,\tau}(f, \mathbf{x}) \wedge \gamma_{s,\tau}(f, \mathbf{x}'')\}$ .

**Proposition 2.** If  $m \geq \left(\frac{b}{t}\right)^2$ , then

$$\begin{aligned} &\Pr_{\mathbf{z}, \mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \wedge \gamma_{s,\tau}(\hat{f}, \mathbf{x}) \wedge \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \wedge P \hat{f}_\phi - P_z \hat{f}_\phi \geq t \right\} \\ &\leq 2 \Pr_{\mathbf{z}, \mathbf{z}', \mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \wedge \gamma_{s,\tau}(\hat{f}, \mathbf{x}) \wedge \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \wedge \left( P_{\mathbf{z}'} \hat{f}_\phi - P_z \hat{f}_\phi \geq \frac{t}{2} \right) \right\}. \quad \blacksquare \end{aligned} \tag{8}$$

<sup>2</sup>The cardinality matches the size of the training sample  $\mathbf{z}$  purely for simplicity.



Now, observe that the probability of interest can be split into the probability of a large deviation happening under a “good” event and the probability of a “bad” event occurring:

$$\begin{aligned}
& \Pr_{\mathbf{z}\mathbf{z}'\mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}) \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \cap \left( P_{\mathbf{z}'} \hat{f}_\phi - P_{\mathbf{z}} \hat{f}_\phi \geq \frac{t}{2} \right) \right\} \\
&= \Pr_{\mathbf{z}\mathbf{z}'\mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}) \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \cap \gamma_{\eta,s,\tau}(\hat{f}, \mathbf{x}') \cap \left( P_{\mathbf{z}'} \hat{f}_\phi - P_{\mathbf{z}} \hat{f}_\phi \geq \frac{t}{2} \right) \right\} \\
&\quad + \Pr_{\mathbf{z}\mathbf{z}'\mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}) \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \cap \overline{\gamma_{\eta,s,\tau}(\hat{f}, \mathbf{x}')} \cap \left( P_{\mathbf{z}'} \hat{f}_\phi - P_{\mathbf{z}} \hat{f}_\phi \geq \frac{t}{2} \right) \right\} \\
&\leq \Pr_{\mathbf{z}\mathbf{z}'} \left\{ \exists f \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(f, \mathbf{x}) \cap \gamma_{\eta,s,\tau}(f, \mathbf{x}') \cap \left( P_{\mathbf{z}'} f_\phi - P_{\mathbf{z}} f_\phi \geq \frac{t}{2} \right) \right\} \\
&\quad + \Pr_{\mathbf{x}'\mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(\hat{f}, \mathbf{x}'') \cap \overline{\gamma_{\eta,s,\tau}(\hat{f}, \mathbf{x}')} \right\}.
\end{aligned}$$

We treat the first probability in the next subsection. To bound the second probability, note that for each choice of  $\mathbf{x}$ ,  $\hat{f}$  is a fixed function. Hence, it is sufficient to select  $\eta$  such that, for *any* fixed function  $f \in \mathcal{F}$ :

$$\Pr_{\mathbf{x}'\mathbf{x}''} \left\{ \gamma_{s,\tau}(f, \mathbf{x}'') \cap \overline{\gamma_{\eta,s,\tau}(f, \mathbf{x}')} \right\} \leq \delta.$$

**Lemma 4 (Unlikely Bad Ghost).** *Let  $f \in \mathcal{F}$  be fixed. If  $\eta = 2 \log \frac{2}{\delta}$ , then*

$$\Pr_{\mathbf{x}'\mathbf{x}''} \left\{ \gamma_{s,\tau}(f, \mathbf{x}'') \cap \overline{\gamma_{\eta,s,\tau}(f, \mathbf{x}')} \right\} \leq \delta.$$

*Proof sketch:* The proof just uses the same standard permutation argument from the proof of the Good Ghost Lemma (Lemma 2). ■

## 5.2 Rademacher bound in the case of the good event

We now bound the probability of a large deviation in the (likely) case of the good event. For readability, let  $\mathcal{F}_{\mu^*,\gamma}(\mathbf{x})$  be shorthand for  $\{f \in \mathcal{F}_{\mu^*} : \gamma_{s,\tau}(f, \mathbf{x})\}$ . Likewise, let  $\mathcal{F}_{\mu^*,\gamma_\eta}(\mathbf{x})$  be shorthand for  $\{f \in \mathcal{F}_{\mu^*} : \gamma_{\eta,s,\tau}(f, \mathbf{x})\}$ . Let  $\sigma_1, \dots, \sigma_m$  be independent Rademacher random variables distributed uniformly on  $\{-1, 1\}$ .

**Lemma 5 (Symmetrization by Random Signs).**

$$\begin{aligned}
& \Pr_{\mathbf{z}\mathbf{z}'} \left\{ \exists f \in \mathcal{F}_{\mu^*} \cap \gamma_{s,\tau}(f, \mathbf{x}) \cap \gamma_{\eta,s,\tau}(f, \mathbf{x}') \cap \left( P_{\mathbf{z}'} f_\phi - P_{\mathbf{z}} f_\phi \geq \frac{t}{2} \right) \right\} \\
&\leq \Pr_{\mathbf{z},\sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*,\gamma}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\} + \Pr_{\mathbf{z},\sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*,\gamma_\eta}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\}.
\end{aligned}$$

*Proof sketch:* The proof uses a standard application of symmetrization by random signs. ■

Note that  $\mathcal{F}_{\mu^*,\gamma}(\mathbf{x})$  is just  $\mathcal{F}_{\mu^*,\gamma_0}(\mathbf{x})$ , and so we need only bound the second term of the last line above for arbitrary  $\eta \in [m]$ . Before bounding this quantity, which for fixed  $\mathbf{x}$  is a conditional Rademacher average (see Appendix D.1 for the fundamentals of Rademacher and Gaussian averages), we need to establish a few results on the Gaussian average of a related function class.

First, note that for any  $D \in \mathcal{D}$ ,  $D$  can be decomposed into a product of two matrices as  $D = US$ , where all  $U \in \mathcal{U} \subset \mathbb{R}^{d \times k}$  satisfy the isometry property  $U^T U = I$  and  $S \in \mathcal{S} := (B_{\mathbb{R}^k})^k$  (Maurer and Pontil, 2008). Fix  $S$ , a linear hypothesis  $w \in \mathcal{W}$ , and a sample  $\mathbf{x}$  of  $m$  points. The subclass of interest will be those functions corresponding to  $U \in \mathcal{U}$  such that  $\|\varphi_{US}(\mathbf{x})\|_0 \leq s$ . It turns out that the Gaussian average of this subclass is well-behaved.

**Theorem 3 (Gaussian Average for Fixed  $S$  and  $w$ ).** Let  $S \in \mathcal{S}$ ,  $s \in [k]$ , and  $\mathbf{x}$  be a fixed  $m$ -sample. Define a particular subclass of  $\mathcal{U}$  as  $\mathcal{U}_{\mathbf{x}} := \{U \in \mathcal{U} : A_s(US, \mathbf{x})\}$ . Then

$$\mathbb{E}_{\gamma} \sup_{U \in \mathcal{U}_{\mathbf{x}}} \frac{2}{m} \sum_{i=1}^m \gamma_i \langle w, \varphi_{US}(x_i) \rangle \leq \frac{4rk\sqrt{2s}}{\mu_{2s}(S)\sqrt{m}}. \quad (9)$$

The proof of this result uses the following lemma that shows how the difference between the feature maps  $\varphi_{US}$  and  $\varphi_{U'S}$  can be characterized by the difference between  $U$  and  $U'$ . Define the  $s$ -restricted 2-norm of  $S$  as  $\|S\|_{2,s} := \sup_{\{t \in \mathbb{R}^n : \|t\|=1, |\text{supp}(t)| \leq s\}} \|St\|_2$ .

**Lemma 6 (Difference Bound For Isometries).** Let  $x \in B_{\mathbb{R}^d}$ . If  $\|\varphi_{US}(x)\|_0 \leq s$  and  $\|\varphi_{U'S}(x)\|_0 \leq s$ , then

$$\|\varphi_{US}(x) - \varphi_{U'S}(x)\| \leq \frac{2\|S\|_{2,2s}}{\mu_{2s}(S)} \|(U'^T - U^T)x\|_2.$$

*Proof sketch:* The proof uses a perturbation analysis of solutions to linearly constrained positive definite quadratic programs (Daniel, 1973), exploiting the sparsity of the optimal solutions to have dependence only on  $\|S\|_{2,2s}$  and  $\mu_{2s}(S)$  rather than  $\|S\|_2$  and  $\mu_k(S)$ . ■

*Proof sketch (of Theorem 3):* The proof controls the Gaussian average of the function class  $\mathcal{U}_{\mathbf{x}}$ , which is *nonlinear* in  $U$ , by the Gaussian average of a second function class that is *linear* in  $U$ . Lemma 6 is critical in establishing a link between the variances of the Gaussian process induced by each function class, after which Slepian's Lemma (Lemma 9) relates the corresponding Gaussian averages. By exploiting the isometry property of  $\mathcal{U}_{\mathbf{x}}$ , the Gaussian average of the second function class can admit a dimension-free bound. ■

The palette for the pivotal art of this section is established.

**Theorem 4 (Rademacher Average of Mostly Good Random Subclasses).**

$$\Pr_{\mathbf{z}, \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \tau_{\eta}}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\} \leq \binom{m}{\eta} \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_3^2/(2b^2)),$$

for

$$t_3 := \frac{t}{4} - L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) - \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}^*\sqrt{m}} - \frac{2b\eta}{m}.$$

*Proof sketch:* We decompose the space of dictionaries  $\mathcal{D}$  via  $\mathcal{U}$  and  $\mathcal{S}$  and split this function class into  $N = \binom{m}{\eta}$  subclasses, each of which has the property that all functions in the class have a common set of  $m - \eta$  good indices. We then approximate each subclass by a product of  $\varepsilon$ -covers of  $\mathcal{W}$  and (a suitably incoherent subset of)  $\mathcal{S}$ . The error from this approximation is low due to the Sparse Coding Stability Theorem (Theorem 1). Fixing a subclass and a representative from each of the two  $\varepsilon$ -covers induces a smaller subclass, and there is a index set of at least  $m - \eta$  points that are sparsely coded for all encoders in this smaller subclass. Theorem 3 then essentially implies a bound on the Rademacher complexity of each smaller subclass. The union bound over  $[N]$  and the two  $\varepsilon$ -covers finishes the result. ■

For the case of  $\eta = 0$ , define

$$t_2 := \frac{t}{4} - L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) - \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}^*\sqrt{m}}.$$

Since  $\mathcal{F}_{\mathbf{r}, \mathbf{x}}$  is equivalent to  $\mathcal{F}_{\mathbf{r}_0, \mathbf{x}}$ , Lemma 5 and Theorem 4 imply that

$$\begin{aligned} & \Pr_{\mathbf{z}, \mathbf{z}'} \left\{ \exists f \in \mathcal{F}_{\mu^*} \quad \mathcal{R}_{s, \tau}(f, \mathbf{x}) \wedge \mathcal{R}_{\eta, s, \tau}(f, \mathbf{x}') \wedge \left( \mathbb{P}_{\mathbf{z}'} f_{\phi} - \mathbb{P}_{\mathbf{z}} f_{\phi} \geq \frac{t}{2} \right) \right\} \\ & \leq \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_2^2/(2b^2)) + \binom{m}{\eta} \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_3^2/(2b^2)) \\ & \leq 2 \binom{m}{\eta} \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_3^2/(2b^2)). \end{aligned}$$

Finally, the full probability (8) can be upper bounded (using  $\eta = 2 \log \frac{2}{\delta}$ ) as:

$$\begin{aligned} & \Pr_{\mathbf{z}, \mathbf{x}''} \left\{ \hat{f} \in \mathcal{F}_{\mu^*} \cap \gamma_{s, \tau}(\hat{f}, \mathbf{x}) \cap \gamma_{s, \tau}(\hat{f}, \mathbf{x}'') \cap |P_{\hat{f}} - P_{\mathbf{z}} \hat{f}| \geq t \right\} \\ & \leq 4 \binom{m}{2 \log \frac{2}{\delta}} \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_3^2/(2b^2)) + 2\delta. \end{aligned}$$

After some elementary manipulations and choosing  $\varepsilon = \frac{1}{m}$ , we nearly have the final learning bound. Let  $\iota'(\lambda, \mu, m, s) := \frac{3}{m} \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{\mu} + \frac{1}{\lambda} + 1 \right)$ .

**Theorem 5.** Let  $\mu_s^*, \mu_{2s}^* > 0$ ,  $s \in [k]$ , and  $\tau \geq \iota'(\lambda, \mu_s^*, m, s)$ . Suppose an algorithm is trained on a labeled sample  $\mathbf{z}$  of  $m$  points and learns hypothesis  $\hat{f}$ . Let  $\mathbf{x}''$  be a second, unlabeled sample of  $m$  points. Suppose that  $\mu_{2s}(\hat{f}) \geq \mu_{2s}^*$ ,  $\mu_s(\hat{f}) \geq \mu_s^*$ ,  $A_s(\hat{f}, \mathbf{x})$ ,  $A_s(\hat{f}, \mathbf{x}'')$ ,  $C_\tau(\hat{f}, \mathbf{x})$ , and  $C_\tau(\hat{f}, \mathbf{x}'')$  all hold. Then with probability at least  $1 - \delta$  over  $\mathbf{z}$  and  $\mathbf{x}''$ :

$$\begin{aligned} (P - P_{\mathbf{z}})\hat{f}_\phi & \leq \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}^*\sqrt{m}} + b\sqrt{\frac{2((k+1)k \log(8m) + k \log \frac{r}{2} + (2 \log m + 3) \log \frac{16}{\delta})}{m}} \\ & \quad + \frac{L}{m} \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{4b}{m} \log \frac{16}{\delta}. \end{aligned}$$

A bound adaptive to the sparsity level and coding margin on  $\mathbf{z}$  and  $\mathbf{x}''$  is now immediate. Recall that  $\nu(\mu) = \min\{1, 2^{\lfloor \log_2 \mu \rfloor}\}$  for  $\mu > 0$ .

**Corollary 2 (Infinite-Dimensional Learning Bound).** Suppose an algorithm is trained on a labeled sample  $\mathbf{z}$  of  $m$  points and learns hypothesis  $\hat{f}$ . Let  $\mathbf{x}''$  be a second, unlabeled sample of  $m$  points. Choose

$$s = \max \left\{ s' \in [k] : A_{s'}(\hat{f}, \mathbf{x}) \cap A_{s'}(\hat{f}, \mathbf{x}'') \right\} \quad \text{and} \quad \tau = \min \left\{ \tau' \in \mathbb{R}_+ : C_{\tau'}(\hat{f}, \mathbf{x}) \cap C_{\tau'}(\hat{f}, \mathbf{x}'') \right\}.$$

Let  $\hat{\mu}_s := \nu(\mu_s(\hat{f}))$  and  $\hat{\mu}_{2s} := \nu(\mu_{2s}(\hat{f}))$ . If  $\tau \geq \iota'(\lambda, \mu_s(\hat{f}), m, s)$  and  $\mu_{2s}(\hat{f}) > 0$ , then with probability at least  $1 - \delta$  over  $\mathbf{z}$  and  $\mathbf{x}''$ :

$$\begin{aligned} (P - P_{\mathbf{z}})\hat{f}_\phi & \leq \frac{2L\sqrt{\pi}rk\sqrt{s}}{\hat{\mu}_{2s}\sqrt{m}} + b\sqrt{\frac{2\left((k^2 + k) \log(8m) + k \log \frac{r}{2} + (2 \log m + 3) \log \frac{44(\log_2(\hat{\mu}_s) \log_2(\hat{\mu}_{2s}))^2 k}{\delta}\right)}{m}} \\ & \quad + \frac{L}{m} \left( \frac{r}{\hat{\mu}_s} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{4b}{m} \log \frac{44(\log_2(\hat{\mu}_s) \log_2(\hat{\mu}_{2s}))^2 k}{\delta}. \quad \blacksquare \end{aligned}$$

## 6 Discussion

We have shown the first generalization error bounds for predictive sparse coding. The results highlight the central role of the stability of the sparse encoder. The presented bounds are data-dependent and exploit properties relating to the training sample and the learned hypothesis. A learning bound appropriate for the overcomplete setting exhibits square root dependence on both the ambient dimension  $d$  and the size of the dictionary  $k$ . In the infinite-dimensional setting, a dimension-free learning bound has linear dependence on  $k$ , square root dependence on  $s$ , and inverse dependence on the  $2s$ -incoherence.

Maurer and Pontil (2008) previously showed the following generalization bound for unsupervised ( $\ell_1$ -regularized) sparse coding:

$$\Pr_{\mathbf{x}} \left\{ \sup_{D \in \mathcal{D}} |P_D f_D - P_{\mathbf{x}} f_D| \geq \frac{k}{\sqrt{m}} \left( \frac{20}{\lambda} + \frac{1}{2} \sqrt{\log(16m/\lambda^2)} \right) + \sqrt{\frac{\log(1/\delta)}{2m}} \right\} \leq \delta.$$

where  $f_D(x) := \min_{z \in \mathbb{R}^k} \|x - Dz\|_2^2 + \lambda \|z\|_1$ . Comparing their result to Corollary 2 and neglecting regularization parameters, the dimension-free bound in the predictive case is larger by a factor of  $\sqrt{s}$ . It is unclear whether the  $\sqrt{s}$  term is avoidable in the predictive setting. At least from our analysis, it appears that the  $\sqrt{s}$  factor is the price of stability.

The data-dependent stability conditions under which our bounds hold may seem restrictive. In the case where the coding margin over the training sample is not uniformly large, or the sparsity level over the training sample is not uniformly small, bounds based on unlikely large deviations from the expectations of these quantities can be established. We did not pursue such bounds here because in predictive sparse coding all training samples empirically do have a low sparsity level and high  $s$ -incoherence. Also, the analysis becomes considerably more involved but we anticipate the learning bounds to be of similar order.

If one entertains a mixture of  $\ell_1$  and  $\ell_2$  norm regularization as in the elastic net (Zou and Hastie, 2005), fall-back guarantees are possible in both scenarios. A considerably simpler, data-independent analysis is possible in the overcomplete setting with a final bound that essentially just trades  $\mu_s(D)$  for the  $\ell_2$ -norm regularization parameter  $\lambda_2$ . In the infinite-dimensional setting, a simpler non-data-dependent analysis using our approach would only attain a bound of the larger order  $\frac{k^{3/2}}{\lambda_2 \sqrt{m}}$ . In conclusion, the presented data-dependent bounds provide motivation for an algorithm to prefer dictionaries for which small subdictionaries are well-conditioned and to additionally encourage large coding margin on the training sample.

## References

- M. Salman Asif and Justin Romberg. On the LASSO and Dantzig selector equivalence. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1–6. IEEE, 2010.
- David M. Bradley and J. Andrew Bagnell. Differentiable sparse coding. *Advances in Neural Information Processing Systems*, 21:113–120, 2009.
- Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.
- James W. Daniel. Stability of the solution of definite quadratic programs. *Mathematical Programming*, 5(1):41–53, 1973. ISSN 0025-5610.
- R. Herbrich and R.C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2002.
- Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 793–800. MIT Press, 2009.
- Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991. ISBN 3540520139.
- Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1033–1040. MIT Press, 2009.
- Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2010.
- Andreas Maurer and Massimiliano Pontil. Generalization bounds for  $K$ -dimensional coding schemes in Hilbert spaces. In *Algorithmic Learning Theory*, pages 79–91. Springer, 2008.

- Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.
- Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
- Shahar Mendelson and Petra Philips. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.
- Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels*. MIT press Cambridge, Mass, 2002.
- John Shawe-Taylor, Peter L., Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *Information Theory, IEEE Transactions on*, 44(5):1926–1940, 1998.
- David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.
- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.
- Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein. The sample complexity of dictionary learning. In *Conference on Learning Theory*, 2011.
- Vladimir Vapnik and Alexey Chervonenkis. Uniform convergence of frequencies of occurrence of events to their probabilities. In *Dokl. Akad. Nauk SSSR*, volume 181, pages 915–918, 1968.
- M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer, London, 2002.
- Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2223–2231. MIT Press, 2009.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

## A Proof of Sparse Coding Stability Theorem

*Proof (of Theorem 1):* Let  $\alpha$  be equal to  $(\varphi_D(x))_{\mathcal{A}}$ , and define the scaled sign vector  $\zeta := \lambda \text{sign}(\alpha)$ . Our strategy will be to show that, for some  $\Delta \in \mathbb{R}^s$ , the optimal perturbed solution  $\varphi_{\tilde{D}}(x)$  satisfies  $(\varphi_{\tilde{D}}(x))_{\mathcal{A}} = \alpha + \Delta$  and  $(\varphi_{\tilde{D}}(x))_{\mathcal{A}^c} = 0$ , where  $\mathcal{A}^c := [k] \setminus \mathcal{A}$ .

From the optimality conditions for the LASSO (e.g. see optimality conditions L1 and L2 of [Asif and Romberg \(2010\)](#)), it is sufficient to find  $\Delta$  such that

$$\begin{aligned} \langle \tilde{D}_j, x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta) \rangle &= \zeta_j \quad \text{if } j \in \mathcal{A}, \\ \left| \langle \tilde{D}_j, x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta) \rangle \right| &< \lambda \quad \text{otherwise.} \end{aligned}$$

We proceed by setting up the linear system and characterizing the solution vector  $\Delta$ :

$$\tilde{D}_{\mathcal{A}}^T(x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta)) = \zeta \quad \xrightarrow{\text{Solve for } \Delta} \quad \Delta = (\tilde{D}_{\mathcal{A}}^T \tilde{D}_{\mathcal{A}})^{-1}(\tilde{D}_{\mathcal{A}}^T(x - \tilde{D}_{\mathcal{A}}\alpha) - \zeta).$$

Since  $\tilde{D} = D + E$  for  $\|E\|_2 \leq \varepsilon$ ,

$$\begin{aligned} \tilde{D}_{\mathcal{A}}^T(x - \tilde{D}_{\mathcal{A}}\alpha) &= (D_{\mathcal{A}} + E_{\mathcal{A}})^T(x - (D_{\mathcal{A}} + E_{\mathcal{A}})\alpha) \\ &= D_{\mathcal{A}}^T(x - D_{\mathcal{A}}\alpha) - D_{\mathcal{A}}^T E_{\mathcal{A}}\alpha + E_{\mathcal{A}}^T(x - (D_{\mathcal{A}} + E_{\mathcal{A}})\alpha) \\ &= \zeta - D_{\mathcal{A}}^T E_{\mathcal{A}}\alpha + E_{\mathcal{A}}^T(x - (D_{\mathcal{A}} + E_{\mathcal{A}})\alpha), \end{aligned}$$

and so the solution for  $\Delta$  can be reformulated as

$$\Delta = (\tilde{D}_{\mathcal{A}}^T \tilde{D}_{\mathcal{A}})^{-1} (-D_{\mathcal{A}}^T E_{\mathcal{A}} \alpha + E_{\mathcal{A}}^T (x - (D_{\mathcal{A}} + E_{\mathcal{A}}) \alpha)).$$

Now,

$$\begin{aligned} \|\Delta\|_2 &\leq \|\tilde{D}_{\mathcal{A}}^T \tilde{D}_{\mathcal{A}}\|_2^{-1} (\|D_{\mathcal{A}}^T E_{\mathcal{A}} \alpha\|_2 + \|E_{\mathcal{A}}^T (x - (D_{\mathcal{A}} + E_{\mathcal{A}}) \alpha)\|_2) \\ &\leq \frac{1}{\mu} \left( \frac{\varepsilon \sqrt{s}}{\lambda} + \varepsilon \right) \\ &= \frac{\varepsilon}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right). \end{aligned}$$

For  $y \in \mathbb{R}^s$ , let  $y_{k \times 1}$  be the extension to  $\mathbb{R}^k$  satisfying  $(y_{k \times 1})_{\mathcal{A}} = y$  and  $(y_{k \times 1})_{\mathcal{A}^c} = 0$ . For  $(\alpha + \Delta)_{k \times 1}$  to be optimal for  $\text{LASSO}(\lambda, \tilde{D}, x)$ ,  $(\alpha + \Delta)_{k \times 1}$  must satisfy the two optimality conditions and  $\Delta$  must be small enough such that sign consistency holds between  $\alpha$  and  $(\alpha + \Delta)$  (i.e.  $\text{sign}(\alpha_j) = \text{sign}(\alpha_j + \Delta_j)$  for all  $j \in [s]$ ).

We first check the optimality conditions. The first optimality condition is equivalent to

$$\langle \tilde{D}_j, x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta) \rangle = \lambda \quad \text{for } j \in \mathcal{A};$$

this condition is satisfied by construction. The second optimality condition is equivalent to

$$\left| \langle \tilde{D}_j, x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta) \rangle \right| < \lambda \quad \text{for } j \notin \mathcal{A}.$$

But for  $j \notin \mathcal{A}$ ,

$$\begin{aligned} \left| \langle \tilde{D}_j, x - \tilde{D}_{\mathcal{A}}(\alpha + \Delta) \rangle \right| &= |\langle D_j + E_j, x - (D_{\mathcal{A}} + E_{\mathcal{A}})(\alpha + \Delta) \rangle| \\ &= |\langle D_j, x - D_{\mathcal{A}} \alpha \rangle - \langle D_j, D_{\mathcal{A}} \Delta \rangle + \langle E_j, x - (D_{\mathcal{A}} + E_{\mathcal{A}})(\alpha + \Delta) \rangle - \langle D_j, E_{\mathcal{A}}(\alpha + \Delta) \rangle| \\ &< \lambda - \tau + \frac{\varepsilon \sqrt{s}}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \varepsilon + \frac{\varepsilon}{\lambda} \\ &= \lambda - \tau + \varepsilon \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{\mu} + \frac{1}{\lambda} + 1 \right), \end{aligned}$$

and so this condition is satisfied provided that

$$\varepsilon \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{\mu} + \frac{1}{\lambda} + 1 \right) \leq \tau.$$

Now, we check sign consistency. Clearly sign consistency holds over  $\mathcal{A}^c$ . It remains to check that it holds over  $\mathcal{A}$ . Observe that

$$\|\Delta\|_{\infty} \leq \|\Delta\|_2 \leq \frac{\varepsilon}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right).$$

Hence, sign consistency holds provided that

$$|\alpha_j| > \varepsilon \left( \frac{1}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) \right).$$

All the above constraints are satisfied if  $\tau$  satisfies

$$\varepsilon \left( \frac{\frac{s}{\lambda} + \sqrt{s}}{\mu} + \frac{1}{\lambda} + 1 \right) \leq \tau. \quad \blacksquare$$

## B Proof of Symmetrization by Ghost Sample Lemma

*Proof (of Lemma 1):* Replace  $\mathcal{F}(\sigma_n)$  from the notation of Mendelson and Philips (2004) with  $\mathcal{F}(\mathbf{z}, \mathbf{x}'')$ . A modified one-sided version of (Mendelson and Philips, 2004, Lemma 2.2) that uses the more favorable Chebyshev-Cantelli inequality implies that, for every  $t > 0$ :

$$\begin{aligned} & \left(1 - \frac{4 \sup_{f \in \mathcal{F}} \text{Var}(\phi \circ f)}{4 \sup_{f \in \mathcal{F}} \text{Var}(\phi \circ f) + mt^2}\right) \Pr_{\mathbf{z}, \mathbf{x}''} \{ \exists f \in \mathcal{F}(\mathbf{z}, \mathbf{x}'') \text{ } P f_\phi - P_{\mathbf{z}} f_\phi \geq t \} \\ & \leq \Pr_{\mathbf{z}, \mathbf{z}', \mathbf{x}''} \left\{ \exists f \in \mathcal{F}(\mathbf{z}, \mathbf{x}'') \text{ } P_{\mathbf{z}'} f_\phi - P_{\mathbf{z}} f_\phi \geq \frac{t}{2} \right\}. \end{aligned}$$

Furthermore, because we assume a bounded loss function with losses trapped in  $[0, b]$ , it follows that  $\sup_{f \in \mathcal{F}} \text{Var}(\phi \circ f) \leq \frac{b^2}{4}$ . The lemma follows since the left hand term of the LHS of the above inequality is at least  $\frac{1}{2}$  whenever  $m \geq \left(\frac{b}{t}\right)^2$ .  $\blacksquare$

## C Proofs for overcomplete setting

*Proof (of the Good Ghost Lemma (Lemma 2)):* By the assumptions of the lemma, we consider  $D$  satisfying  $\mu_s(D) \geq \mu$  and  $A_s(D, \mathbf{x})$ .

We want to guarantee with high probability that for  $D$  all but  $\eta$  points of the ghost sample (1) are coded at sparsity level  $s$ , (2) have encodings whose non-zero-valued coordinates have absolute value greater than  $\tau_3(D, \mathbf{x})$ , and (3) have encodings whose zero-valued coordinates have absolute correlation of the corresponding atom with the residual (i.e.  $|\langle D_i, x_i - D\varphi_D(x_i) \rangle|$ ) less than  $\lambda - \tau_3(D, \mathbf{x})$ .

The latter two inequality conditions are the *raison d'être* of  $\psi$ . Consider the class of threshold functions

$$\mathcal{F}_D^{\text{stab}} := \{f_{D, \tau}^{\text{stab}} | \tau \in \mathbb{R}_+\}$$

defined as

$$f_{D, \tau}^{\text{stab}}(x) := \begin{cases} 1; & \text{if } \max_{j \in [k]} \psi_j(D, X) < \lambda - \tau, \\ 0; & \text{otherwise.} \end{cases}$$

Since the VC dimension of the one-dimensional threshold functions is 1, it follows that the  $\text{VC}(\mathcal{F}_D^{\text{stab}}) = 1$ .

Consider a minimum-cardinality  $\left(\frac{\tau_3(D, \mathbf{x})\mu}{\frac{s+\mu}{\lambda} + \sqrt{s+\mu}}\right)$ -proper cover  $\mathcal{D}'$  of  $\mathcal{D}_\mu$ . Let  $D'$  be a candidate element of  $\mathcal{D}'$  satisfying  $\|D - D'\|_2 \leq \frac{\tau_3(D, \mathbf{x})\mu}{\frac{s+\mu}{\lambda} + \sqrt{s+\mu}}$ . Then the Sparse Coding Stability Theorem (Theorem 1) implies  $A_s(D', \mathbf{x})$  and  $C_{\tau_{3/2}(D, \mathbf{x})}(D', \mathbf{x})$ . First, we ensure that most points from the ghost sample satisfy  $\max_{j \in [k]} \psi_j(D', \cdot) < \lambda - \tau_{3/2}(D, \mathbf{x})$ . By using the VC dimension of  $\mathcal{F}_D^{\text{stab}}$  and the standard permutation argument of (Vapnik and Chervonenkis, 1968, Proof of Theorem 2), it follows that for a single, *fixed* element of  $\mathcal{D}'$ , with probability at least  $1 - \delta$  at most  $\log(2m + 1) + \log \frac{1}{\delta}$  points from a ghost sample will violate the inequality. Hence, by the bound on the proper covering numbers provided by Proposition 3 (see Appendix E), we can guarantee for all candidate members  $D' \in \mathcal{D}'$  that with probability  $1 - \frac{\delta}{2}$  at most

$$\begin{aligned} & dk \log \left( \frac{8 \left( \frac{s+\mu}{\lambda} + \sqrt{s+\mu} \right)}{\tau_3(D, \mathbf{x})\mu} \right) + \log(2m + 1) + \log \frac{2}{\delta} \\ & \leq dk \log \frac{96s}{\lambda\mu\tau(D, \mathbf{x})} + \log(2m + 1) + \log \frac{2}{\delta} \end{aligned}$$

points from the ghost sample violate the inequality (in the above, we use the generally loose bound  $\mu \leq s$  and assume without loss of generality<sup>3</sup> that  $\lambda \leq 1$ ).

<sup>3</sup>If  $\lambda > 1$ , then all sparse codes will equal the zero vector.



It also is necessary to guarantee with high probability that most points of the ghost sample are coded sparsely. Since  $s$  is fixed, for a single element of  $\mathcal{D}'$  we need only bound the probability that a single function

$$f_{D,s}^{\text{spar}}(x) := \begin{cases} 1; & \text{if } \|\varphi_D(x)\|_0 \leq s \\ 0; & \text{otherwise} \end{cases}$$

takes the value 1 for all points in the training sample but takes the value of 0 for some number of points in the ghost sample. Indeed a standard permutation argument shows that  $\log \frac{2}{\delta}$  or more points of the ghost sample will violate the sparsity condition with probability at most  $\frac{\delta}{2}$ .

Thus, for arbitrary  $D' \in \mathcal{D}'$  satisfying the conditions of the lemma, with probability  $1 - \delta$  at most

$$\eta(m, d, k, \varepsilon, \delta) := 2dk \log \frac{96s}{\lambda \mu \tau(D, \mathbf{x})} + \log(2m + 1) + 2 \log \frac{2}{\delta}$$

points from the ghost sample violate  $A_s(D', \mathbf{x}')$  or  $C_{\tau_{3/2}(D, \mathbf{x})}(D', \mathbf{x}')$ .

Now, consider the at least  $m - \eta(m, d, k, \varepsilon, \delta)$  points in the ghost sample that satisfy both  $A_s(D', \cdot)$  and  $C_{\tau_{3/2}(f, \mathbf{x})}(D', \cdot)$ . Since  $\|D' - D\|_2 \leq \frac{\tau_{3/2}(D, \mathbf{x})\mu}{\frac{\tau_{3/2}(D, \mathbf{x})\mu}{s+\mu} + \sqrt{s+\mu}}$ , the Sparse Coding Stability Theorem (Theorem 1) implies that these points satisfy  $A_s(D, \cdot)$  and  $C_{\tau_3(D, \mathbf{x})}(D, \cdot)$ . ■

*Proof (of Lemma 3):* First, note that  $J \cap \bar{Z}$  is a subset of

$$R := \left\{ \mathbf{z}\mathbf{z}' : \begin{array}{l} \exists f \in \mathcal{F}_\mu \quad \mathcal{R}_{s,t}(f, \mathbf{x}) \\ \wedge \left( \nexists \tilde{\mathbf{x}} \subset \mathbf{x}', \|\tilde{\mathbf{x}}\| > \eta, \forall \mathbf{x} \in \tilde{\mathbf{x}} \quad \overline{\mathcal{R}_{s,\tau_3(f, \mathbf{x})}(f, \mathbf{x})} \right) \\ \wedge (P_{\mathbf{z}'} f_\phi - P_{\mathbf{z}} f_\phi > t/2) \end{array} \right\}.$$

To see this, let **bad ghost** be shorthand for the condition of their being more than  $\eta$  points in the ghost sample that violate  $\mathcal{R}_{s,\tau_3(f, \mathbf{x})}(f, \cdot)$  (i.e., that are “bad”), let **good ghost** be true if and only if **bad ghost** is false, and let  $P$  be shorthand for the condition  $P_{\mathbf{z}'} f_\phi - P_{\mathbf{z}} f_\phi > t/2$ . Suppose that  $\mathbf{z}\mathbf{z}' \in J \cap \bar{Z}$ . Hence,  $\exists f \in \mathcal{F}_\mu$  such that  $\mathcal{R}_{s,t}(f, \mathbf{x})$  and  $P$ , and furthermore,  $\nexists \tilde{\mathbf{x}} \in \mathcal{F}_\mu$  such that  $\mathcal{R}_{s,t}(f, \mathbf{x})$  and **bad ghost**. Consider all  $f \in \mathcal{F}_\mu$  such that  $\mathcal{R}_{s,t}(f, \mathbf{x})$  and  $P$ . Suppose any one of these satisfies **bad ghost**. Then  $\exists f \in \mathcal{F}_\mu$  such that  $\mathcal{R}_{s,t}(f, \mathbf{x})$  and **bad ghost**, which implies that  $\mathbf{z}\mathbf{z}' \notin J \cap \bar{Z}$ . Hence, if  $\mathbf{z}\mathbf{z}' \in J \cap \bar{Z}$  then  $\exists f \in \mathcal{F}_\mu$  such that  $\mathcal{R}_{s,t}(f, \mathbf{x})$ ,  $P$ , and **good ghost**.

Now, choose a  $f = (D, w) \in \mathcal{F}_\mu$  satisfying  $\mathcal{R}_{s,t}(f, \mathbf{x})$  and **good ghost**. Thus, if  $\|D - D'\|_2 \leq \varepsilon$  and  $\|w - w'\|_2 \leq \varepsilon$ , then for all but  $\eta$  of the points in the ghost sample (and for all points of the original sample) we are guaranteed that

$$\begin{aligned} & |\langle w, \varphi_D(x_i) \rangle - \langle w', \varphi_{D'}(x_i) \rangle| \\ & \leq |\langle w - w', \varphi_D(x_i) \rangle| + |\langle w', \varphi_D(x_i) - \varphi_{D'}(x_i) \rangle| \\ & \leq \frac{\varepsilon}{\lambda} + r \frac{\varepsilon}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) \quad (\text{Sparse Coding Stability Theorem (Theorem 1)}) \\ & = \varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) =: \beta. \end{aligned}$$

For the rest of the ghost sample, there is a coarse guarantee that  $\|\varphi_D(x_i) - \varphi_{D'}(x_i)\|_2 \leq \frac{2}{\lambda}$ . Hence, on the original sample

$$\frac{1}{m} \sum_{i=1}^m |\phi(y_i, \langle w, \varphi_D(x_i) \rangle) - \phi(y_i, \langle w', \varphi_{D'}(x_i) \rangle)| \leq L\beta,$$

where the loss  $\phi$  is  $L$ -Lipschitz. On the ghost sample

$$\begin{aligned}
& \frac{1}{m} \sum_{i=1}^m |\phi(y'_i, \langle w, \varphi_D(x'_i) \rangle) - \phi(y'_i, \langle w', \varphi_{D'}(x'_i) \rangle)| \\
& \leq \frac{L}{m} \left( \sum_{i \text{ good}} |\langle w, \varphi_D(x_i) \rangle - \langle w', \varphi_{D'}(x_i) \rangle| \right) + \frac{1}{m} \sum_{i \text{ bad}} |\phi(y'_i, \langle w, \varphi_D(x'_i) \rangle) - \phi(y'_i, \langle w', \varphi_{D'}(x'_i) \rangle)| \\
& \leq L\beta + \frac{\eta}{m} \min \left\{ \frac{L(\varepsilon + 2r)}{\lambda}, b \right\} \\
& \leq L\beta + \frac{b\eta}{m}.
\end{aligned}$$

The subclass of interest is

$$\tilde{\mathcal{F}}(\mathbf{x}, \mathbf{x}') := \left\{ f \in \mathcal{F}_\mu : \mathcal{I}_{s,t}(f, \mathbf{x}) \wedge \left( \# \tilde{\mathbf{x}} \subset \mathbf{x}', |\tilde{\mathbf{x}}| > \eta, \forall x \in \tilde{\mathbf{x}} \overline{\mathcal{I}_{s,t_3}(f, \mathbf{x})}(f, x) \right) \right\}.$$

It is sufficient to consider the  $\varepsilon$ -neighborhood of  $\tilde{\mathcal{F}}(\mathbf{x}, \mathbf{x}')$  defined such that if  $f = (D, w) \in \tilde{\mathcal{F}}(\mathbf{x}, \mathbf{x}')$ , then all  $f' = (D', w')$  satisfying  $\|D - D'\|_2 \leq \varepsilon$  and  $\|w - w'\|_2 \leq \varepsilon$  are in the  $\varepsilon$ -neighborhood.

Let  $\mathcal{F}_\varepsilon = \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$ , where  $\mathcal{D}_\varepsilon$  is a minimum-cardinality proper  $\varepsilon$ -cover of  $\mathcal{D}_\mu$  and  $\mathcal{W}$  is a minimum-cardinality  $\varepsilon$ -cover of  $\mathcal{W}$ . Note that the  $\varepsilon$ -neighborhood will contain at least one element of  $\mathcal{F}_\varepsilon$ . Hence, it is sufficient to prove the large deviation bound for all of  $\mathcal{F}_\varepsilon$  and to then consider the maximum difference between an element of  $\tilde{\mathcal{F}}(\mathbf{x}, \mathbf{x}')$  and its closest representative in  $\mathcal{F}_\varepsilon$  (which of course is  $O(\varepsilon)$ ).

Now, for each  $f = (D, w) \in \mathcal{F}_\mu$  satisfying  $\mathcal{I}_{s,t}(f, \mathbf{x})$  and **goodghost**, there is a  $D' \in \mathcal{D}_\varepsilon$  such that  $\|D - D'\|_2 \leq \varepsilon$  and a  $w' \in \mathcal{W}_\varepsilon$  satisfying  $\|w - w'\|_2 \leq \varepsilon$ ; hence, the difference between the losses of  $f$  and  $f'$  on the double sample will be at most  $2L\beta + \frac{b\eta}{m}$ .

Let  $\mathbf{v}$  be the absolute deviation between the loss of  $f$  on the original sample versus its loss on the ghost sample. Then the absolute deviation between the loss of  $f' = (D', w')$  on the original sample and the loss of  $f'$  on the ghost sample is at least

$$\mathbf{v} - \left( 2L\beta + \frac{b\eta}{m} \right).$$

Hence, if  $\mathbf{v} > t/2$ , then the absolute deviation between the loss of  $f'$  on the original sample and the loss of  $f'$  on the ghost sample must be at least

$$t/2 - \left( 2L\beta + \frac{b\eta}{m} \right).$$

Let  $f_{D',w'}$  be the hypothesis corresponding to  $(D', w')$ . It therefore is sufficient to bound the probability that

$$\Pr_{\mathbf{z}, \mathbf{z}'} \left\{ \exists f = (D', w') \in \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon \text{ } \mathbb{P}_{\mathbf{z}'} f_\phi - \mathbb{P}_{\mathbf{z}} f_\phi > t/2 - \left( 2L\beta + \frac{b\eta}{m} \right) \right\},$$

since  $R$  implies the above event.

Now, let  $\omega := t/2 - \left( 2L\beta + \frac{b\eta}{m} \right)$ .

We first handle the case of a fixed  $f = (D', w') \in \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$ . Just apply Hoeffding's inequality for the random variable  $\phi(y, f(x_i)) - \phi(y, f(x'_i))$  (range in  $[-b, b]$ ) to yield

$$\Pr_{\mathbf{z}, \mathbf{z}'} \{ \mathbb{P}_{\mathbf{z}'} f_\phi - \mathbb{P}_{\mathbf{z}} f_\phi > \omega \} \leq \exp(-m\omega^2/(2b^2)).$$

Apply Proposition 4 to extend this bound over all of  $\mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon$  via the union bound, yielding

$$\begin{aligned}
& \Pr_{\mathbf{z}, \mathbf{z}'} \{ \exists f = (D', w') \in \mathcal{D}_\varepsilon \times \mathcal{W}_\varepsilon \text{ } \mathbb{P}_{\mathbf{z}'} f_\phi - \mathbb{P}_{\mathbf{z}} f_\phi > \omega \} \\
& \leq \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\omega^2/(2b^2)).
\end{aligned}$$

Thus,

$$\Pr(J \cap \bar{Z}) \leq \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)). \quad \blacksquare$$

*Proof (of Theorem 2):* Proposition 1 and Lemmas 2 and 3 imply that

$$\begin{aligned} & \Pr_{\mathbf{z}} \{ \exists f \in \mathcal{F}_{\mu} \ \gamma_{s,t}(f, \mathbf{x}) \wedge (P f_{\phi} - P_{\mathbf{z}} f_{\phi} > t) \} \\ & \leq 2 \left( \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \delta \right) \end{aligned}$$

Equivalently,

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \ \gamma_{s,t}(f, \mathbf{x}) \wedge \left( P f_{\phi} - P_{\mathbf{z}} f_{\phi} > 2 \left( \varpi + 2L\beta + \frac{b\eta(m, d, k, \varepsilon, \delta)}{m} \right) \right) \right\} \\ & \leq 2 \left( \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \delta \right) \end{aligned}$$

Now, expand  $\beta$  and  $\eta$  and replace  $\delta$  with  $\delta/4$ :

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \ \gamma_{s,t}(f, \mathbf{x}) \wedge P f_{\phi} - P_{\mathbf{z}} f_{\phi} \right. \\ & \quad \left. > 2 \left( \varpi + 2L\varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{b(2dk \log \frac{96s}{\lambda\mu\tau(\bar{f}, \mathbf{x})} + \log(2m+1) + 2 \log \frac{8}{\delta})}{m} \right) \right\} \\ & \leq 2 \left( \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \frac{\delta}{2} \right). \end{aligned}$$

Choose  $\frac{\delta}{4} = \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2))$ , yielding

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \ \gamma_{s,t}(f, \mathbf{x}) \right. \\ & \quad \wedge P f_{\phi} - P_{\mathbf{z}} f_{\phi} > 2 \left( \varpi + 2L\varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) \right) \\ & \quad \left. + 2 \frac{b(2dk \log \frac{96s}{\lambda\mu\tau(\bar{f}, \mathbf{x})} + \log(2m+1) + 2(d+1)k \log \frac{\varepsilon}{8(r/2)^{1/(d+1)}} + \frac{m\varpi^2}{b^2} + 2 \log 2)}{m} \right\} \\ & \leq 4 \cdot \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)). \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \ \gamma_{s,t}(f, \mathbf{x}) \right. \\ & \quad \wedge P f_{\phi} - P_{\mathbf{z}} f_{\phi} > 2 \left( \varpi + 2L\varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) \right) \\ & \quad \left. + 2 \frac{b(2dk \log \frac{96s}{\lambda\mu\tau(\bar{f}, \mathbf{x})} - 2(d+1)k \log \frac{8}{\varepsilon} + 2k \log \frac{2}{r} + \log(2m+1) + \frac{m\varpi^2}{b^2} + 2 \log 2)}{m} \right\} \\ & \leq 4 \cdot \left( \frac{8(r/2)^{1/(d+1)}}{\varepsilon} \right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)). \end{aligned}$$

Let  $\delta$  (a new variable, not related to the previous incarnation of  $\delta$ ) be equal to the upper bound, and solve for  $\omega$ , yielding:

$$\omega = b \sqrt{\frac{2((d+1)k \log \frac{8}{\varepsilon} + k \log \frac{r}{2} + \log \frac{4}{\delta})}{m}}$$

and hence

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \quad \gamma_{s,l}(f, \mathbf{x}) \right. \\ & \quad \wedge \quad \mathbb{P} f_{\Phi} - \mathbb{P}_{\mathbf{z}} f_{\Phi} > 2b \sqrt{\frac{2((d+1)k \log \frac{8}{\varepsilon} + k \log \frac{r}{2} + \log \frac{4}{\delta})}{m}} \\ & \quad \left. + 2 \left( 2L\varepsilon \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{b(2dk \log \frac{96s}{\lambda \mu \tau(f, \mathbf{x})} + \log(2m+1) + 2 \log \frac{8}{\delta})}{m} \right) \right\} \\ & \leq \delta. \end{aligned}$$

If we set  $\varepsilon = \frac{1}{m}$ , then provided that  $m > \frac{1}{3\tau} \left( \frac{\frac{\varepsilon}{\lambda} + \sqrt{s}}{\mu} + 1 \right)$ :

$$\begin{aligned} & \Pr_{\mathbf{z}} \left\{ \exists f \in \mathcal{F}_{\mu} \quad \gamma_{s,l}(f, \mathbf{x}) \right. \\ & \quad \wedge \quad \mathbb{P} f_{\Phi} - \mathbb{P}_{\mathbf{z}} f_{\Phi} > 2b \sqrt{\frac{2((d+1)k \log(8m) + k \log \frac{r}{2} + \log \frac{4}{\delta})}{m}} \\ & \quad \left. + \frac{4L}{m} \left( \frac{r}{\mu} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{2b}{m} \left( 2dk \log \frac{96s}{\lambda \mu \tau(f, \mathbf{x})} + \log(2m+1) + 2 \log \frac{8}{\delta} \right) \right\} \\ & \leq \delta. \quad \blacksquare \end{aligned}$$

## D Infinite-dimensional setting

### D.1 Rademacher and Gaussian averages and related results

Let  $\sigma_1, \dots, \sigma_m$  be independent Rademacher random variables distributed uniformly on  $\{-1, 1\}$ , and let  $\gamma_1, \dots, \gamma_m$  be independent Gaussian random variables distributed as  $\mathcal{N}(0, 1)$ .

Given a sample of  $m$  points  $\mathbf{x}$ , define the conditional Rademacher and Gaussian averages of a function class as

$$\mathcal{R}_{m|\mathbf{x}}(\mathcal{F}) = \frac{2}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i) \quad \text{and} \quad \mathcal{G}_{m|\mathbf{x}}(\mathcal{F}) = \frac{2}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^m \gamma_i f(x_i).$$

respectively.

From [Meir and Zhang \(2003, Theorem 7\)](#), the loss-composed conditional Rademacher average of a function class  $\mathcal{F}$  is bounded by the scaled conditional Rademacher average:

**Lemma 7 (Rademacher Loss Comparison Lemma).** *For every function class  $\mathcal{F}$ , sample of  $m$  points  $\mathbf{x}$ , and  $\phi$  which is  $L$ -Lipschitz continuous in its second argument:*

$$\mathcal{R}_{m|\mathbf{z}}(\phi \circ \mathcal{F}) \leq L \mathcal{R}_{m|\mathbf{x}}(\mathcal{F}).$$

Additionally, from [Ledoux and Talagrand \(1991\)](#), a brief argument following Lemma 4.5), the conditional Rademacher average of a function class  $\mathcal{F}$  is bounded up to a constant by the conditional Gaussian average of  $\mathcal{F}$ :

**Lemma 8 (Rademacher-Gaussian Average Comparison Lemma).** *For every function class  $\mathcal{F}$  and sample of  $m$  points  $\mathbf{x}$ :*

$$\mathcal{R}_{m|\mathbf{x}}(\mathcal{F}) \leq \sqrt{\frac{\pi}{2}} \mathcal{G}_{m|\mathbf{x}}(\mathcal{F}).$$

The next relation is due to [Slepian \(1962\)](#).

**Lemma 9 (Slepian's Lemma).** *Let  $\Omega$  and  $\Gamma$  be mean zero, separable Gaussian processes indexed by a set  $T$  such that  $\mathbb{E}(\Omega_{t_1} - \Omega_{t_2})^2 \leq \mathbb{E}(\Gamma_{t_1} - \Gamma_{t_2})^2$  for all  $t_1, t_2 \in T$ . Then  $\mathbb{E} \sup_{t \in T} \Omega_t \leq \mathbb{E} \sup_{t \in T} \Gamma_t$ .*

Slepian's Lemma essentially says that if the variance of one Gaussian process is bounded by the variance of another, then, in expectation, the  $\sup$  of the first is bounded by the  $\sup$  of the second.

We also will make use of McDiarmid's inequality ([McDiarmid, 1989](#)).

**Theorem 6 (McDiarmid's Inequality).** *Let  $X_1, \dots, X_m$  be random variables drawn iid according to a probability measure  $\mu$  over a space  $\mathcal{X}$ . Suppose that a function  $f : \mathcal{X}^m \rightarrow \mathbb{R}$  satisfies*

$$\sup_{x_1, \dots, x_m, x'_i \in \mathcal{X}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c_i$$

for any  $i \in [m]$ . Then

$$\Pr_{X_1, \dots, X_n} \{f(X_1, \dots, X_n) - \mathbb{E} f(X_1, \dots, X_n) \geq t\} \leq \exp \left( -2t^2 / \sum_{i=1}^m c_i^2 \right).$$

## D.2 Proofs

*Proof (of Lemma 4):* By definition,  $\Pr_{\mathbf{x}', \mathbf{x}''} \left\{ \gamma_{s, \tau}(f, \mathbf{x}'') \wedge \overline{\gamma_{\eta, s, \tau}(f, \mathbf{x}')} \right\}$  is equal to

$$\Pr_{\mathbf{x}', \mathbf{x}''} \left\{ A_s(f, \mathbf{x}'') \wedge C_\tau(f, \mathbf{x}'') \wedge \exists \tilde{\mathbf{x}} \subset \mathbf{x}' \ |\tilde{\mathbf{x}}| > \eta \wedge \forall x \in \tilde{\mathbf{x}} \ \overline{A_s(f, x)} \vee \overline{C_\tau(f, x)} \right\}.$$

From the permutation argument, if no point in  $\mathbf{x}''$  violates  $A_s(f, \cdot)$ , then the probability that over  $\log \frac{2}{\delta}$  points of  $\mathbf{x}'$  will violate  $A_s(f, \cdot)$  is at most  $\frac{\delta}{2}$ ; in addition, if no point of  $\mathbf{x}''$  violates  $C_\tau(f, \cdot)$ , then the probability that over  $\log \frac{2}{\delta}$  points of  $\mathbf{x}'$  will violate  $C_\tau(f, \cdot)$  is at most  $\frac{\delta}{2}$ . Thus, the probability that more than  $\eta = 2 \log \frac{2}{\delta}$  points of  $\mathbf{x}'$  violate either  $A_s(f, \cdot)$  or  $C_\tau(f, \cdot)$  is at most  $\delta$ . ■

*Proof (of Lemma 5):* From the definitions of  $\gamma_{s, \tau}$  and  $\gamma_{\eta, s, \tau}$ :

$$\begin{aligned} & \Pr_{\mathbf{z}, \mathbf{z}'} \left\{ \exists f \in \mathcal{F}_{\mu^*} \ \gamma_{s, \tau}(f, \mathbf{x}) \wedge \gamma_{\eta, s, \tau}(f, \mathbf{x}') \wedge \left( P_{\mathbf{z}'} f_\Phi - P_{\mathbf{z}} f_\Phi \geq \frac{t}{2} \right) \right\} \\ &= \Pr_{\mathbf{z}, \mathbf{z}'} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, r}(\mathbf{x}) \cap \mathcal{F}_{\mu^*, r_\eta}(\mathbf{x}')} \frac{1}{m} \sum_{i=1}^m (\phi(y'_i, f(x'_i)) - \phi(y_i, f(x_i))) \geq \frac{t}{2} \right\}. \end{aligned}$$

Now, a routine application of symmetrization by random signs yields

$$\begin{aligned}
& \Pr_{\mathbf{z}\mathbf{z}'} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \gamma}(\mathbf{x}) \cap \mathcal{F}_{\mu^*, \gamma_\eta}(\mathbf{x}')} \frac{1}{m} \sum_{i=1}^m (\phi(y'_i, f(x'_i)) - \phi(y_i, f(x_i))) \geq \frac{t}{2} \right\} \\
&= \Pr_{\mathbf{z}\mathbf{z}', \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \gamma}(\mathbf{x}) \cap \mathcal{F}_{\mu^*, \gamma_\eta}(\mathbf{x}')} \frac{1}{m} \sum_{i=1}^m \sigma_i (\phi(y'_i, f(x'_i)) - \phi(y_i, f(x_i))) \geq \frac{t}{2} \right\} \\
&\leq \Pr_{\mathbf{z}\mathbf{z}', \sigma} \left\{ \left( \sup_{f \in \mathcal{F}_{\mu^*, \gamma}(\mathbf{x}) \cap \mathcal{F}_{\mu^*, \gamma_\eta}(\mathbf{x}')} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right) \vee \left( \sup_{f \in \mathcal{F}_{\mu^*, \gamma}(\mathbf{x}) \cap \mathcal{F}_{\mu^*, \gamma_\eta}(\mathbf{x}')} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y'_i, f(x'_i)) \geq \frac{t}{4} \right) \right\} \\
&\leq \Pr_{\mathbf{z}, \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \gamma}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\} + \Pr_{\mathbf{z}, \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \gamma_\eta}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\}. \quad \blacksquare
\end{aligned}$$

*Proof (of Lemma 6):* By definition,  $\varphi_{US}(\mathbf{x}) = \arg \min_{\mathbf{z} \in \mathbb{R}^k} \|\mathbf{x} - US\mathbf{v}\|_2 + \lambda \|\mathbf{z}\|_1$ . First, note the equivalences  $\arg \min_{\mathbf{z} \in \mathbb{R}^k} \|\mathbf{x} - US\mathbf{z}\|_2 = \arg \min_{\mathbf{z} \in \mathbb{R}^k} \|U^T \mathbf{x} - U^T US\mathbf{z}\|_2 = \arg \min_{\mathbf{z} \in \mathbb{R}^k} \|U^T \mathbf{x} - S\mathbf{z}\|_2$ , where the first equality follows because any  $\mathbf{z} \in \text{Ker}(U)$  (i.e., in the complement of the range space of  $U$ ) will be orthogonal to  $US\mathbf{z}$ , for any  $\mathbf{z}$ ; hence, it is sufficient to approximate the projection of  $\mathbf{x}$  onto the range space of  $U$ .

Thus,  $\varphi_{US}(\mathbf{x}) = \arg \min_{\mathbf{z} \in \mathbb{R}^k} \|U^T \mathbf{x} - S\mathbf{z}\|_2^2 + \lambda \|\mathbf{z}\|_1$ . Our approach is to use the well-known reformulation as a quadratic program with linear constraints:

$$\begin{aligned}
& \underset{\bar{\mathbf{z}} \in \mathbb{R}^{3k}}{\text{minimize}} \quad Q_U(\bar{\mathbf{z}}) := \bar{\mathbf{z}}^T \begin{pmatrix} S^T S & 0_{k \times 2k} \\ 0_{2k \times k} & 0_{2k \times 2k} \end{pmatrix} \bar{\mathbf{z}} - \bar{\mathbf{z}}^T \begin{pmatrix} 2S^T U^T \\ 0_{2k \times d} \end{pmatrix} \mathbf{x} + \lambda(0_k^T \mathbf{1}_{2k}^T) \bar{\mathbf{z}} \\
& \text{subject to} \quad \mathbf{z}^+ \geq 0_k \quad \mathbf{z}^- \geq 0_k \quad \mathbf{z} - \mathbf{z}^+ + \mathbf{z}^- = 0_k,
\end{aligned}$$

where  $\bar{\mathbf{z}} := \bar{\mathbf{z}} := (\mathbf{z}^T \mathbf{z}^{+T} \mathbf{z}^{-T})^T$ .

For optimal solutions  $\bar{\mathbf{z}}_* := \begin{pmatrix} \mathbf{z}_* \\ \mathbf{z}_*^+ \\ \mathbf{z}_*^- \end{pmatrix}$  and  $\bar{\mathbf{t}}_* := \begin{pmatrix} \mathbf{t}_* \\ \mathbf{t}_*^+ \\ \mathbf{t}_*^- \end{pmatrix}$  of  $Q_U$  and  $Q_{U'}$  respectively, from [Daniel \(1973\)](#),

we have

$$(\bar{\mathbf{u}} - \bar{\mathbf{z}}_*)^T \nabla Q_U(\bar{\mathbf{z}}_*) \geq 0 \quad (10)$$

$$(\bar{\mathbf{u}} - \bar{\mathbf{t}}_*)^T \nabla Q_{U'}(\bar{\mathbf{t}}_*) \geq 0 \quad (11)$$

for all  $\bar{\mathbf{u}} \in \mathbb{R}^{3k}$ . Setting  $\bar{\mathbf{u}}$  to  $\bar{\mathbf{t}}_*$  in (10) and  $\bar{\mathbf{u}}$  to  $\bar{\mathbf{z}}_*$  in (11) and adding (10) and (11) yields

$$(\bar{\mathbf{t}}_* - \bar{\mathbf{z}}_*)^T (\nabla Q_U(\bar{\mathbf{z}}_*) - \nabla Q_{U'}(\bar{\mathbf{t}}_*)) \geq 0,$$

which is equivalent to

$$(\bar{\mathbf{t}}_* - \bar{\mathbf{z}}_*)^T (\nabla Q_{U'}(\bar{\mathbf{t}}_*) - \nabla Q_{U'}(\bar{\mathbf{z}}_*)) \leq (\bar{\mathbf{t}}_* - \bar{\mathbf{z}}_*)^T (\nabla Q_U(\bar{\mathbf{z}}_*) - \nabla Q_{U'}(\bar{\mathbf{z}}_*)). \quad (12)$$

Here,  $\nabla Q_U(\mathbf{z}) = \begin{pmatrix} S^T S & 0_{k \times 2k} \\ 0_{2k \times k} & 0_{2k \times 2k} \end{pmatrix} \mathbf{z} - \begin{pmatrix} 2S^T U^T \\ 0_{2k \times d} \end{pmatrix} \mathbf{x} + \lambda \begin{pmatrix} 0_k \\ \mathbf{1}_{2k} \end{pmatrix}$ . After plugging in the expansions of  $\nabla Q_U$  and  $\nabla Q_{U'}$  and incurring cancellations from the zeros, (12) becomes

$$\begin{aligned}
& (\mathbf{t}_* - \mathbf{z}_*)^T (S^T S \mathbf{t}_* - 2S^T U'^T \mathbf{x} - S^T S \mathbf{z}_* + 2S^T U'^T \mathbf{x}) \leq (\mathbf{t}_* - \mathbf{z}_*)^T (S^T S \mathbf{z}_* - 2S^T U^T \mathbf{x} - S^T S \mathbf{z}_* + 2S^T U'^T \mathbf{x}) \\
& \quad \Downarrow \\
& (\mathbf{t}_* - \mathbf{z}_*)^T S^T S (\mathbf{t}_* - \mathbf{z}_*) \leq 2(\mathbf{t}_* - \mathbf{z}_*)^T S^T (U'^T - U^T) \mathbf{x}.
\end{aligned}$$

If  $\mathbf{t}_*$  and  $\mathbf{z}_*$  both have sparsity level at most  $s$ , then wherever we typically would consider the operator norm  $\|S\|_2 := \sup_{\|t\|_1=1} \|St\|_2$ , we instead need only consider the  $2s$ -restricted operator norm  $\|S\|_{2,2s}$ .

Note that  $(t_* - z_*)^T S^T S(t_* - z_*) \geq \mu_{2s}(S) \|t_* - z_*\|_2^2$ , which implies that

$$\|t_* - z_*\|_2^2 \leq \frac{2}{\mu_{2s}(S)} \|t_* - z_*\| \|S\|_{2,2s} \|(U'^T - U^T)x\| \Rightarrow \|t_* - z_*\|_2 \leq \frac{2\|S\|_{2,2s}}{\mu_{2s}(S)} \|(U'^T - U^T)x\|_2. \quad \blacksquare$$

*Proof (of Theorem 3):* Define a Gaussian process  $\Omega$ , indexed by  $U$ , by  $\Omega_U := \sum_{i=1}^m \gamma_i \langle w, \varphi_{US}(x_i) \rangle$ . Our goal is to apply Slepian's Lemma to bound the expectation of the supremum of  $\Omega$ , which depends on  $\varphi_{US}$ , by the expectation of the supremum of a Gaussian process  $\Gamma$  which depends only on  $U$ .

$$\begin{aligned} \mathbb{E}_\gamma (\Omega_U - \Omega_{U'})^2 &= \mathbb{E}_\gamma \left( \sum_{i=1}^m \gamma_i \langle w, \varphi_{US}(x_i) \rangle - \sum_{i=1}^m \gamma_i \langle w, \varphi_{U'S}(x_i) \rangle \right)^2 \\ &= \sum_{i=1}^m (\langle w, \varphi_{US}(x_i) - \varphi_{U'S}(x_i) \rangle)^2 \\ &\leq r^2 \sum_{i=1}^m \|\varphi_{US}(x_i) - \varphi_{U'S}(x_i)\|^2 \end{aligned} \quad (13)$$

Applying the result from Lemma 6, we have

$$\begin{aligned} \mathbb{E}_\gamma (\Omega_U - \Omega_{U'})^2 &\leq r^2 \sum_{i=1}^m \|\varphi_{US}(x_i) - \varphi_{U'S}(x_i)\|^2 \\ &\leq \left( \frac{2r\|S\|_{2,2s}}{\mu_{2s}(S)} \right)^2 \sum_{i=1}^m \|(U'^T - U^T)x_i\|_2^2 \\ &= \left( \frac{2r\|S\|_{2,2s}}{\mu_{2s}(S)} \right)^2 \sum_{i=1}^m \sum_{j=1}^k (\langle U'e_j, x_i \rangle - \langle Ue_j, x_i \rangle)^2 \\ &= \left( \frac{2r\|S\|_{2,2s}}{\mu_{2s}(S)} \right)^2 \mathbb{E}_\gamma \left( \left( \sum_{i=1}^m \sum_{j=1}^k \gamma_{ij} \langle U'e_j, x_i \rangle \right) - \left( \sum_{i=1}^m \sum_{j=1}^k \gamma_{ij} \langle Ue_j, x_i \rangle \right) \right)^2 \\ &= \mathbb{E}_\gamma (\Gamma_U - \Gamma_{U'})^2 \end{aligned}$$

for

$$\Gamma_U := \frac{2r\|S\|_{2,2s}}{\mu_{2s}(S)} \sum_{i=1}^m \sum_{j=1}^k \gamma_{ij} \langle Ue_j, x_i \rangle.$$



By Slepian's Lemma,  $E_\gamma \sup_U \Omega_U \leq E_\gamma \sup_U \Gamma_U$ . It remains to bound  $E_\gamma \sup_U \Gamma_U$ :

$$\begin{aligned}
\frac{\mu_{2s}(S)}{2r\|S\|_{2,2s}} E_\gamma \sup_U \Gamma_U &= E_\gamma \sup_U \sum_{i=1}^m \sum_{j=1}^k \gamma_{ij} \langle Ue_j, x_i \rangle \\
&= E_\gamma \sup_U \sum_{j=1}^k \langle Ue_j, \sum_{i=1}^m \gamma_{ij} x_i \rangle \\
&\leq E_\gamma \sup_U \sum_{j=1}^k \|Ue_j\| \left\| \sum_{i=1}^m \gamma_{ij} x_i \right\| \\
&= k E_\gamma \left\| \sum_{i=1}^m \gamma_{i1} x_i \right\| \\
&\leq k \sqrt{E_\gamma \left\| \sum_{i=1}^m \gamma_{i1} x_i \right\|^2} \\
&= k \sqrt{E_\gamma \left\langle \sum_{i=1}^m \gamma_{i1} x_i, \sum_{i=1}^m \gamma_{i1} x_i \right\rangle} = k \sqrt{\sum_{i=1}^m \|x_i\|^2} \leq k \sqrt{m}.
\end{aligned}$$

Hence,

$$E_\gamma \sup_{U \in \mathcal{U}} \frac{2}{m} \sum_{i=1}^m \gamma_i \langle w, \varphi_{US}(x_i) \rangle \leq \frac{4r\|S\|_{2,2s}k}{\mu_{2s}(S)\sqrt{m}} \leq \frac{4rk\sqrt{2s}}{\mu_{2s}(S)\sqrt{m}},$$

where we used the fact that  $\|S\|_{2,2s} \leq \sqrt{2s}$  (see Lemma 10 in Appendix D.2 for a proof). ■

*Proof (of Theorem 4):* In the course of the proof, we will use the following factorization of the space of dictionaries  $\mathcal{D}$  that was introduced by [Maurer and Pontil \(2008\)](#). Each dictionary  $D \in \mathcal{D}$  will be factorized as  $D = US$ , where  $S \in \mathcal{S} := (B_{\mathbb{R}^k})^k$  and

$$U \in \mathcal{U} := \{U' \in \mathbb{R}^{d \times k} : \|U'x\|_2 = \|x\|_2 \text{ for } x \in \mathbb{R}^k\}.$$

$\mathcal{U}$  is the set of semi-orthogonal matrices in  $\mathbb{R}^{d \times k}$ ; these matrices are left isometries (i.e. they satisfy  $U^T U = I$ ). Let  $\mathcal{S}_\varepsilon$  be a minimum-cardinality proper  $\varepsilon$ -cover (in operator norm) of  $\{S \in \mathcal{S} : \mu_s(S) \geq \mu_s^*, \mu_{2s}(S) \geq \mu_{2s}^*\}$ , the set of suitably incoherent elements of  $\mathcal{S}$ .

Now, onward to the proof. The goal is to bound

$$\Pr_{\mathbf{x}, \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\}.$$

We say an index  $i$  is good if and only if  $\tau_{s, \tau}(f, x_i)$ . An index is bad if and only if it is not good. Consider a fixed sample  $\mathbf{z}$  and the occurrence of a set of  $m - \eta$  good indices; each of the remaining indices can be either good or bad. There are  $N := \binom{m}{\eta}$  ways to choose this set of indices. Partition  $\mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x})$  into  $N$  subclasses via  $\mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x}) = \bigcup_{j \in [N]} \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x})$  where, for all functions in a given subclass, a specific set of  $m - \eta$  indices is guaranteed to be good. More formally, we can choose distinct good index sets  $\Gamma_1, \dots, \Gamma_N$ , each of size  $m - \eta$ , such that for each  $j$ :

$$\Gamma_j \subset \bigcap_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x})} \{i \in [m] : \tau_{s, \tau}(f, x_i)\}.$$

Clearly,

$$\sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x})} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) = \max_{j \in [N]} \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x})} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)).$$

For each  $j \in [N]$ , define the  $\varepsilon$ -neighborhood of  $\mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x})$  defined as

$$\bar{\mathcal{F}}_{\mu^*, \tau_\eta}^j(\mathbf{x}) := \left\{ f = (US', w') : \|S - S'\| \leq \varepsilon, \|w - w'\| \leq \varepsilon, S \in \mathcal{S}, w \in \mathcal{W}, (US, w) \in \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x}) \right\}.$$

Also, let  $\mathcal{W}_\varepsilon$  be a minimum-cardinality  $\varepsilon$ -cover of  $\mathcal{W}$  and define a particular epsilon-cover of  $\mathcal{F}$ :

$$\mathcal{F}_\varepsilon := \{f = (US', w') \in \mathcal{F} : U \in \mathcal{U}, S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon\}.$$

Take the intersection  $\bar{\mathcal{F}}_{\mu^*, \tau_\eta}^j(\mathbf{x}) \cap \mathcal{F}_\varepsilon$ , a disjoint union of subclasses equal to

$$\bigcup_{S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon} \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})$$

for

$$\mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x}) := \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x}) \cap \{f \in \mathcal{F} : f = (US', w') : U \in \mathcal{U}\}.$$

Now, for each  $j \in [N]$  and arbitrary  $\sigma \in \{-1, 1\}^m$ , we compare

$$\sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^j(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \quad \text{and} \quad \max_{S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon} \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)).$$

Without loss of generality, choose  $j = 1$  and take  $\Gamma_1 = [m - \eta]$ . If  $f \in \mathcal{F}_{\mu^*, \tau_\eta}^1(\mathbf{x})$ , it follows that there exists  $f' \in \bigcup_{S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon} \mathcal{F}_{\mu^*, \tau_\eta}^{1, S', w'}(\mathbf{x})$  such that

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m \sigma_i |\phi(y_i, \langle w, \varphi_D(x_i) \rangle) - \phi(y_i, \langle w', \varphi_{D'}(x_i) \rangle)| \\ & \leq \frac{L}{m} \left( \sum_{i=1}^{m-\eta} \sigma_i |\langle w, \varphi_D(x_i) \rangle - \langle w', \varphi_{D'}(x_i) \rangle| \right) + \frac{1}{m} \sum_{i=m-\eta+1}^m \sigma_i |\phi(y_i, \langle w, \varphi_D(x_i) \rangle) - \phi(y_i, \langle w', \varphi_{D'}(x_i) \rangle)| \\ & \leq L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{b\eta}{m}, \end{aligned}$$

where the last line is due to the Sparse Coding Stability Theorem (Theorem 1).

Therefore, for any  $\sigma \in \{-1, 1\}^m$  it holds that

$$\begin{aligned} & \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \\ & \leq \max_{j \in [N]} \max_{S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon} \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) + L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) + \frac{b\eta}{m}. \end{aligned}$$

Hence, for  $\mathbf{z}$  fixed

$$\begin{aligned} & \Pr_\sigma \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\} \\ & \leq \Pr_\sigma \left\{ \max_{j \in [N]} \max_{S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon} \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} - L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) - \frac{b\eta}{m} \right\} \\ & \leq N \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \max_{\substack{j \in [N] \\ S' \in \mathcal{S}_\varepsilon, w' \in \mathcal{W}_\varepsilon}} \Pr_\sigma \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} - L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) - \frac{b\eta}{m} \right\}. \end{aligned}$$

Now, from McDiarmid's inequality, for any fixed  $j \in [N]$ ,  $S' \in \mathcal{S}_\varepsilon$  and  $w' \in \mathcal{W}_\varepsilon$ :

$$\Pr_\sigma \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) > \mathbb{E}_\sigma \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) + t_1 \right\} \leq \exp(-mt_1^2/(2b^2)).$$

Now, we bound

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}_{\mu^*, \tau_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)).$$

Without loss of generality, again take  $j = 1$  and  $\Gamma_1 = [m - \eta]$ . Then

$$\begin{aligned}
& \mathbb{E}_\sigma \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \\
&= \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^{m-\eta} \sigma_i \phi(y_i, f(x_i)) + \frac{1}{m} \sum_{i=m-\eta+1}^m \sigma_i \phi(y_i, f(x_i)) \right\} \\
&\leq \mathbb{E}_{\sigma_1, \dots, \sigma_{m-\eta}} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^{m-\eta} \sigma_i \phi(y_i, f(x_i)) \right\} + \mathbb{E}_{\sigma_{m-\eta+1}, \dots, \sigma_m} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=m-\eta+1}^m \sigma_i \phi(y_i, f(x_i)) \right\} \\
&\leq \mathbb{E}_{\sigma_1, \dots, \sigma_{m-\eta}} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^{m-\eta} \sigma_i \phi(y_i, f(x_i)) \right\} + \frac{b\eta}{m}.
\end{aligned}$$

Now, Theorem 3, the Rademacher Loss Comparison Lemma (Lemma 7), and the Rademacher-Gaussian Average Comparison Lemma (Lemma 8) imply that

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{\substack{U \in \mathcal{U} \\ A_s(U', \mathbf{x})}} \frac{1}{m} \sum_{i=1}^{m-\eta} \sigma_i \phi(y_i, \langle w, \varphi_{US}(x_i) \rangle) &\leq \frac{\sqrt{m-\eta}}{m} \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}(S)} \\
&\leq \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}(S)\sqrt{m}},
\end{aligned}$$

and hence

$$\Pr_\sigma \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(f(x_i)) > \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}^* \sqrt{m}} + \frac{b\eta}{m} + t_1 \right\} \leq \exp(-mt_1^2/(2b^2)).$$

Combining this bound with the fact that the bound is independent of the draw of  $\mathbf{z}$  and applying Proposition 4 (with  $d = k$ ) to extend the bound over all choices of  $j$ ,  $S'$ , and  $w'$  results in

$$\Pr_{\mathbf{z}, \sigma} \left\{ \sup_{f \in \mathcal{F}_{\mu^*, \Gamma_\eta}^{j, S', w'}(\mathbf{x})} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(y_i, f(x_i)) \geq \frac{t}{4} \right\} \leq \binom{m}{\eta} \left( \frac{8(r/2)^{1/(k+1)}}{\varepsilon} \right)^{(k+1)k} \exp(-mt_3^2/(2b^2)),$$

for

$$t_3 := \frac{t}{4} - L\varepsilon \left( \frac{r}{\mu_s^*} \left( \frac{\sqrt{s}}{\lambda} + 1 \right) + \frac{1}{\lambda} \right) - \frac{2L\sqrt{\pi}rk\sqrt{s}}{\mu_{2s}^* \sqrt{m}} - \frac{2b\eta}{m}.$$

■

**Lemma 10.** If  $S \in (B_{\mathbb{R}^k})^k$ , then  $\|S\|_{2,2s} \leq \sqrt{2s}$ .

*Proof:* Define  $S_\Lambda$  as the submatrix of  $S$  that selects the columns indexed by  $\Lambda$ . Similarly, for  $t \in \mathbb{R}^k$  define the coordinate projection  $t_\Lambda$  of  $t$ .

$$\begin{aligned}
& \sup_{\{t: \|t\|=1, |\text{supp}(t)| \leq 2s\}} \|St\|_2 \\
&= \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \|S_\Lambda t_\Lambda\|_2 \\
&= \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \left\| \sum_{\omega \in \Lambda} t_\omega S_\omega \right\|_2 \\
&\leq \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \sum_{\omega \in \Lambda} |t_\omega| \|S_\omega\|_2 \\
&\leq \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \sum_{\omega \in \Lambda} |t_\omega| \\
&\leq \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \|t_\Lambda\|_1 \\
&\leq \min_{\{\Lambda \subset [k]: |\Lambda| \leq 2s\}} \sup_{\{t: \|t\|=1, \text{supp}(t) \subset \Lambda\}} \sqrt{2s} \|t_\Lambda\|_2 \\
&= \sqrt{2s}.
\end{aligned}$$

■

## E Covering numbers

Cucker and Smale (2002, Chapter I, Proposition 5) state that, for a Banach space  $E$  of dimension  $d$ , the  $\varepsilon$ -covering numbers of the radius  $r$  ball of  $E$  are bounded as  $\mathcal{N}(rB_E, \varepsilon) \leq (4r/\varepsilon)^d$ .

For spaces of dictionaries obeying some deterministic property, such as  $\mathcal{D}_\mu = \{D \in \mathcal{D} : \mu_s(D) \geq \mu\}$ , one must be careful to use a *proper*  $\varepsilon$ -cover so that the representative elements of the cover also obey the desired property: A proper cover is more restricted than a cover in that a proper cover must be a subset of the set being covered (rather than simply being a subset of the ambient Banach space). That is, if  $A$  is a proper cover of a subset  $T$  of a Banach space  $E$ , then  $A \subset T$ . For a cover, we need only  $A \subset E$ . The following bound relates proper covering numbers to covering numbers (a simple proof can be found in (Vidyasagar, 2002, Lemma 2.1)): If  $E$  is a Banach space and  $T \subset E$  is a bounded subset, then  $\mathcal{N}(E, \varepsilon, T) \leq \mathcal{N}_{\text{proper}}(E, \varepsilon/2, T)$ .

Let  $d, k \in \mathbb{N}$ . Define  $E_\mu := \{E \in (B_{\mathbb{R}^d})^k : \mu_s(D) \geq \mu\}$  and  $\mathcal{W} := rB_{\mathbb{R}^d}$ . The following bounds derive directly from the above.

**Proposition 3.** The proper  $\varepsilon$ -covering number of  $E_\mu$  is bounded by

$$\left(\frac{8}{\varepsilon}\right)^{dk}.$$

■

**Proposition 4.** The product of the proper  $\varepsilon$ -covering number of  $E_\mu$  and the  $\varepsilon$ -covering number of  $\mathcal{W}$  is bounded by

$$\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k}.$$

■